DETERMINING THE PSYCHOMETRIC PROPERTIES OF THE RETRIEVAL-INDUCED FORGETTING PROCEDURE

A Thesis Submitted to the College of Graduate Studies and Research in Partial Fulfillment of the Requirements for the Degree of Master of Arts in the Department of Educational Psychology University of Saskatchewan Saskatoon

By

JENNIFER L. BRIERE

PERMISSION TO USE

In presenting this thesis/dissertation in partial fulfillment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis/dissertation in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis/dissertation work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis/dissertation or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis/dissertation.

Requests for permission to copy or to make other uses of materials in this thesis/dissertation in whole or part should be addressed to:

Head of the Department of Educational Psychology and Special Education
University of Saskatchewan
Saskatoon, Saskatchewan S7N 0X1
Canada

ABSTRACT

Repeatedly retrieving information from memory has been shown to induce forgetting of related, un-retrieved information below baseline, an effect termed *retrieval-induced forgetting* (RIF; Anderson, Bjork, & Bjork, 1994). In the current research, *stability* and *alternate-forms reliability* estimates of RIF scores were evaluated through correlations of five RIF tasks using two sets of equated category-word pairs and one set of facts in sentence format. *Convergent* and *discriminant validity* estimates were evaluated through correlation of RIF scores and scores on the *Cognitive Failures Questionnaire* (CFQ) and the *Social Desirability Scale–17* (SDS-17), respectively. Analysis indicated that although RIF was obtained on all four tasks, stability reliability was obtained for only the sets of materials that participants completed twice, with no evidence for alternate forms reliability. Stability reliability for the category-word pair RIF task that participants completed twice, two-weeks apart, accounted for 17.6% of the variance in scores, r (50) = .42, p = .003. The facts RIF task was completed again approximately one month following the initial administration and stability reliability was also obtained using these materials, r (18) = .51, p = .032, accounting for 27% of the variance in scores. Evidence of discriminant validity was found through non-significant correlations between the RIF tasks and scores on the SDS-17, however evidence of convergent validity was not obtained when correlating CFQ and RIF scores. The results suggest that variations in the degree of forgetting observed from RIF tasks may largely depend on the type of materials used. Implications for theory and research regarding RIF are discussed.

ACKNOWLEDGEMENTS

Great accomplishments are rarely made alone and completing this research is no different. Throughout my studies I have had many supporters, all of whom have made my navigation through this process easier. I owe many thanks to each and every one of them as they have helped me achieve my goals and shaped who I am.

I owe sincere and earnest thanks to my Mom and Dad, Sandy and Alan Briere, for their never-ending support and encouragement. They have been, and continue to be, my biggest fans. Mom and Dad, I would never be me without you – thank you.

I am truly indebted to my supervisors, Dr. Tammy Marche and Dr. Laurie Hellsten. Without their guidance I would never have been in reach of my goals. I can only aspire to be as amazing as they are. I could not have found a more caring and supportive mentor than Dr. Marche – I am forever indebted to you.

I am obliged to my family and friends who always listened and uplifted me. Your words of wisdom and support reminded me that the tribulations of academia were not insurmountable.

Last, but by far the least, I extend deep thanks to my boyfriend, Jon Berger. Words cannot express how deeply I appreciate his love, support in helping me achieve my dreams and his unending devotion to me. Forever my champion, forever my friend, you Jon, could not be more important to me. Thank you for who you are and who you make me be.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CFQ: Cognitive Failures Questionnaire

DOF: Degree of Forgetting

DRM: Deese (1959) Roediger and McDermott (1995) False Memory Paradigm

GRE: Graduate Record Examination

IQ: Intelligence Quotient

RIF: Retrieval-Induced Forgetting

Rp+: Details used during RIF tasks that receive retrieval-practice

Rp-: Details used during RIF tasks that are from the retrieval-practiced category but do not receive any practice

NRp: Details used during RIF tasks that are from the no retrieval-practice baseline category

SDS – 17: Social Desirability Scale – 17

1. CHAPTER 1: INTRODUCTION

1.1. Determining the Psychometric Properties of Retrieval-Induced Forgetting

Psychological constructs, such as individuals' skills and abilities, are not directly measurable in the same way as physical characteristics and abilities. It is quite easy to obtain a measurement of an individual's height or weight, but it is not so easy to obtain a measure of that individual's level of depression or self-esteem. Instead of taking a physical measurement where there is a known, absolute zero with equal increments between points on the scale as with height and weight measurements, the measurement of psychological constructs must be reduced to observations of behaviours that are theoretically related to the construct of interest, or *latent variable* (Crocker & Algina, 1986; DeVellis, 2003). For example, assigning a proficiency score to an individual based on his or her knowledge of a certain topic cannot be obtained by simply asking the individual what he or she knows about the topic. Instead, that individual's knowledge of the topic is assessed through the use of test questions that have been specifically designed to sample from the content of knowledge that an individual who is proficient in the topic should know (Cronbach, 1951). It is the individual's performance on the test that is being measured and used to calculate that individual's proficiency score, which is then interpreted and used in some manner by the user/administrator (Messick, 1989). Thus, the degree of confidence in the appropriate use of scores can only be as strong as the measurement tool used to assess the construct. How then, does one decide whether or not a specific measure is "strong" enough to warrant its use and whether to be confident in the application and use of scores? *Classical test (true score) theory* would typically be relied upon which assumes that every psychological or educational assessment is made up of both a measure of individuals' true (real) score for the latent variable, as well as statistical or measurement error (i.e., *Observed score = True score +*

Error; Allen & Yen, 1979; DeVellis, 2003; Kline, 2005; Traub, 1997). Using this framework, one would then examine the available psychometric properties of the measure, such as *reliability* and *validity* estimates, to provide statistical guidance towards informed decisions regarding the appropriate use of scores (Crocker & Algina, 1986). The term *reliability* refers to the consistency, stability and/or precision of scores obtained on a measure either during one administration or across time intervals (Allen & Yen, 1979; Cronbach, 1951; DeVellis, 2003; Gall, Gall, & Borg, 2007; Heffner, 2004) while *validity* refers to the appropriateness, meaningfulness, and application of test scores as well as the usefulness of judgments made from test scores (Allen & Yen, 1979; Gall et al., 2007; Heffner, 2004; Messick, 1989). Most psychometrics are evaluated using simple correlation or Pearson's *r* (Allen & Yen, 1979; Crocker & Algina, 1986; DeVellis, 2003) which yields a statistic that reflects the direction and strength of a relationship between two scores. The correlation coefficient provides a highly interpretable numerical value that represents the degree to which change in one variable is met with similar change in the other variable, or the degree to which the variables *covary* (Field, 2009).

Regardless of the importance of understanding the psychometrics of measures, and the ease in which they can be calculated, there are instances when there is a lack of evidence available to inform users about the appropriate interpretation and application of specific scores, which could lead to potentially disastrous outcomes (e.g., not treating a severely depressed client). It is quite unfortunate when researchers and practitioners interpret obtained scores as a measure of some latent variable when psychometrics regarding that measure are unavailable – conclusions drawn based on those scores cannot be made with much confidence as individuals' true score on the construct would not be consistently predicted. Each form of the various types

of reliability and validity estimates provides test users with different information regarding participants' true scores for the latent variable and the associated amount of measurement error. For example, reliability evidence can inform users about the stability of scores across time, individuals and/or materials (Allen & Yen, 1979; Cronbach, 1951; DeVellis, 2003; Gall et al., 2007; Heffner, 2004) while validity evidence can demonstrate the strength of relationships between scores obtained on measures that are theoretically related or theoretically dissimilar (Allen & Yen, 1979; Gall et al., 2007; Heffner, 2004; Messick, 1989). Gaining the most comprehensive understanding possible regarding the psychometrics of measures that one intends to use will likely lead to more appropriate interpretation and application of the scores, which in turn leads to a greater degree of confidence in score use.

The overall goal of the current research is to inform researchers and practitioners about the validity and reliability of scores obtained through a procedure that is considered robust in the literature and to subsequently guide users to more confident interpretations and applications of the scores. The purpose of this research was to illustrate how reliability and validity evidence can be statistically examined after a measure has been used as an individual difference in research rather than during measure development and to subsequently increase confidence in score use. In addition, the research addresses a gap in the literature by providing psychometric evidence where it is greatly needed, and informs current theory and interpretation regarding the construct of interest, an effect termed *retrieval-induced forgetting* (RIF; Anderson, Bjork & Bjork, 1994).

During a RIF procedure (e.g., Anderson et al., 1994), participants first study a number of items, one at a time, that are grouped under semantically related category cues (e.g., fruit – orange, vegetable – carrot, fruit – apple, vegetable – beans, fruit – strawberry, vegetable – onion,

fruit – pear, vegetable – peas). Once participants have studied all of the items, they are given an opportunity to practice retrieving certain items from memory three times each using a fill-in-the-blank task (e.g., fruit – or____, fruit – ap____) that is referred to as *retrieval-practice*. Following a brief distractor task, each category cue is provided to participants one at a time and participants are asked to write down all of the items that they remember being paired with that category during the initial study trial. Participants' final recall data demonstrate both a *practice effect* and an *inhibition effect* that are collectively termed *RIF effects*. The practice effect is illustrated by participants recalling significantly more items that underwent retrieval-practice (termed Rp+ items) when compared to items that did not receive retrieval-practice but were from the practiced category (termed Rp- items) as well as when compared to the no retrieval-practice baseline category (termed NRp items). More interestingly however, is the significant reduction in recall of the Rp- items when compared to the NRp baseline. This reduced recall is believed to reflect inhibition, or temporary forgetting, of the Rp- items (Anderson et al., 1994). In the preceding example, apple and orange would be deemed Rp+ items because they underwent retrieval-practice; strawberry and pear would be deemed Rp- items as they are from the practiced category, but were left out of the retrieval-practice task, and finally all items from the vegetable category would be deemed NRp items as they did not receive any retrieval-practice and are from a different category (i.e., no manipulation, or baseline condition). The act of repeatedly retrieving the Rp+ items (e.g., apple, orange) from memory is argued to create mental competition with Rp- items (e.g., strawberry, pear; Anderson, 2003; Anderson et al., 1994). To facilitate accurate retrieval at final test and perhaps reduce interference (Anderson, 2003), inhibition of Rp- items occurs which drives down recall of these items below the NRp baseline (vegetable) at final test. This pattern of results has been investigated and obtained across a wide

variety of materials (e.g., word lists, Anderson et al., 1994; facts, Macrae & MacLeod, 1999; social cognition, Macrae & MacLeod, 1999; autobiographical memory, Barnier, Hung & Conway, 2004; eyewitness memory, Shaw, Bjork & Handal, 1995; visuospatial memory, Ciranni & Shimamura, 1999) and population samples (e.g., children, Ford, Keating & Patel, 2004; adults, Migueles & García-Bajos, 2007; clinical populations, Nestor et al., 2005) leading researchers to conclude that RIF effects are quite robust. With such a prolific research literature on RIF effects, and researchers' interpretation of RIF scores as a measure of individual difference in unintentional forgetting (e.g., Aslan & Bäuml, 2011; Johansson, Aslan, Bäuml, Gable & Mecklinger, 2007; Marche, Briere & von Baeyer, 2011), it is surprising that the psychometric properties of the scores produced through the procedure have not yet been evaluated. Without knowing how stable RIF scores are, researchers may draw conclusions and make decisions based on scores that may not be consistently tapping true score variance. In other words, the scores obtained and interpreted during any one administration of the measure may not be stable across time, across individuals, within individuals or any combination of the above – strong conclusions cannot be made with confidence as the psychometrics of the scores remains unknown.

Following a list of definitions of important terms used in the current thesis that closes this chapter, the remaining chapters will first briefly review the nature and importance of establishing psychometric properties of measures and will be followed by a discussion of the theoretical framework (*classical test*, or *true-score theory*) that allows psychometric properties to be assessed. The chapter will then move into a review of the different types of psychometric properties that classical test theory affords as well as a discussion of simple correlation, the statistical procedure that permits the evaluation of psychometrics. An example of evaluating the psychometrics of a consistent pattern of results obtained through a manipulation that is argued to

measure an individual difference in a latent variable will then be provided. A gap in the research literature will then be identified where there is an absence of psychometric evidence available regarding a measure in wide use (i.e., RIF). A brief review of the current study that is designed to address this research gap will then be provided followed by the study hypotheses. Chapters 3 and 4 will then provide the reader with a description of the study methodology and results, respectively. The final Discussion chapter provides an interpretation of the results of the study, as well as conclusions, limitations and potential future directions for research in the area.

1.1.1. Definitions of Important Terms

Alternate forms reliability estimates. Estimates that compare scores obtained on an alternate form of the same measure that purports to tap the same (or similar) underlying construct as the measure being evaluated (Allen & Yen, DeVellis, 2003).

Carry-over (practice) effects. Changes in performance on a measure due to previous experience with the same test (Allen & Yen, 1979).

Classical test theory. A theory of educational and psychological measurement. When using classical test theory it is assumed that every measurement is composed of an individual's *true score* of the latent variable (i.e., pure measure of the latent variable) as well as measurement *error* (i.e., random and/or systematic error; Allen & Yen, 1979; Crocker & Algina, 1986; DeVellis, 2003).

Convergent validity. The extent to which scores on a measure share a high, medium, or low relationship with scores obtained on a different measure that is intended to asses the same (or similar) construct (Messick, 1989; 1995).

Coefficient of determination. The squared result of Pearson's r (i.e., r^2) which describes the amount of variance in scores on one measure that is shared by the other measure

(e.g., $r = .24$, $r^2 = .058$, thus the two measures share 5.8% of the variance; Crocker & Algina, 1986; Field, 2009).

Coefficient of stability. The correlation between two administrations of the same test to the same individuals; also referred to as *test-retest reliability*, or *stability reliability* (Allen & Yen, 1979; Crocker & Algina, 1986).

Coefficient of stability and equivalence. The correlation between scores obtained from the same individual on two measures that are matched to each other in every possible way; the impact of carry-over effects is reduced in comparison to the coefficient of stability (Crocker & Algina, 1986; Cronbach, 1951).

Cognitive inhibition. Refers to an individual's ability to supress previously activated cognitive processes or representations (Aslan & Bäuml, 2011; MacLeod, 2002; Wilson & Kipp, 1998).

Concurrent validity. A type of *criterion-related validity* that indicates the degree to which scores obtained on a measure estimate individuals' performance on some criterion measure (Allen & Yen, 1979; Messick, 1989).

Construct validity. The extent to which a measure actually measures what it intends to (Allen & Yen, 1979; DeVellis, 2003; Gall et al., 2007; Heffner, 2004; Messick, 1989).

Content validity. Refers to the degree to which a measure comprehensively assesses the underlying construct of interest; often referred to as *face validity* (Allen & Yen, 1979; DeVellis, 2003; Messick, 1989, 1995).

Criterion related validity. The degree to which scores obtained on an educational or psychological measure relate to one or more outcome criteria (Allen & Yen, 1979; Messick, 1989, 1995). There are two types of criterion related validity: *predictive validity,* and *concurrent*

validity.

Degree of forgetting score. A continuous score that reflects the degree of impairment, or forgetting, that results from using a retrieval-induced forgetting procedure (Anderson et al., 1994).

Discriminant validity. The extent to which scores on a measure *do not* share a relationship with scores obtained on a theoretically unrelated measure (Messick, 1995).

Essentially tau (τ) equivalent. Two halves of a measure that have the same true score for each half of the test but have unequal error variances (Allen & Yen, 1979).

Face validity. Refers to the degree to which a measure comprehensively assesses the underlying construct of interest; often referred to as *content validity* (Allen & Yen, 1979; DeVellis, 2003; Messick, 1989, 1995).

Internal consistency reliability. The degree to which items on a measure correlate with one another, or the degree to which a measure is *homogeneous* (Crocker & Algina, 1986; Cronbach, 1951; DeVellis, 2003).

Inter-rater/observer reliability. The degree to which different raters' or observers' scores on the same measure agree, or are replicable (reliable; Cohen, 1960; Crocker & Algina, 1986).

Latent variable. The underlying construct that a test intends to measure; latent variables are not directly measured but are inferred based on observation of other variables (DeVellis, 2003).

Measurement. In psychological and educational assessment, measurement refers to observations of behaviour that result in a score intended to reflect a certain amount of a latent variable in the individual being observed (Crocker & Algina, 1986).

NRp (no retrieval-practice) details. Details used during retrieval-induced forgetting tasks that are from the no retrieval-practice baseline category. NRp details do not receive any manipulation (Anderson et al., 1994).

Parallel forms reliability. The degree to which scores obtained on two parallel tests (i.e., identical alphas, means and variances) correlate with one another (Crocker & Algina, 1986; DeVellis, 2003).

Pearson's correlation coefficient (*r*). A standardized measure of the strength of the relationship between two variables; sometimes referred to as *Pearson's product-moment correlation coefficient* with values ranging from -1 to +1 (Field, 2009).

Predictive validity. A type of criterion-related validity that reflects the extent to which scores obtained on a measure predict future behaviour on some criterion measure (Allen & Yen, 1979; Messick, 1989; e.g., predicting graduate school performance based on scores obtained on the Graduate Records Examination).

Psychometrics. Theory and/or methods of mental measurement (Allen & Yen, 1979).

Random error. Influences on scores obtained from a measure that are unaccounted for and affect some, but not other, examinees' scores (Allen & Yen, 1979; DeVellis, 2003).

Reliability. The consistency, stability and/or precision of scores obtained on a measure either during one administration or across time intervals (Allen & Yen, 1979; Cronbach, 1951; DeVellis, 2003; Gall et al., 2007; Heffner, 2004; Messick, 1989).

Retrieval-induced forgetting (RIF). A procedure that results in unintentional forgetting of a subset of related information that undergoes extra practice (termed *retrieval-practice*) once all information has been studied once. The procedure progresses along four steps: (a) *initial study* (all information is studied once), (b) *retrieval-practice* (a subset of the related information

is repeatedly retrieved through a fill in the blanks-like task, (c) distractor task (unrelated filler activity occurs for a few minutes, and (d) final recall (cued recall for all information occurs).

Rp+ details. Details used during retrieval-induced forgetting tasks that receive extra practice by retrieving the information from memory using a fill in the blanks-like task.

Rp- details. Details used during retrieval-induced forgetting tasks that do not receive extra practice but are related to the details that received retrieval-practice.

Split-halves reliability. An estimate of internal consistency reliability obtained from splitting a measure into two parallel, or essentially τ equivalent, halves (Allen & Yen, 1979).

Stability reliability. An estimate of the degree to which two administrations of the same test to the same individual(s) correlated with one another; also referred to as *test-retest reliability*, or the *coefficient of stability* (Allen & Yen, 1979; Crocker & Algina, 1986).

Systematic error. Error that influences all examinee's scores in a systematic way (i.e., systematically raises or lowers obtained scores for all examinees; Allen & Yen, 1979; DeVellis, 2003).

Test-retest reliability. An estimate of the degree to which two (or more) administrations of the same test to the same group of examinees correlate with one another; also referred to as *stability reliability*, or the *coefficient of stability* (Allen & Yen, 1979; Crocker & Algina, 1986).

True score (*T*). An individual's *real* or *pure* measure of the latent variable that is free from error (Allen & Yen, 1979; Crocker & Algina, 1986; DeVellis, 2003).

Validity. An overall judgement regarding the degree to which empirical evidence and theory support adequate and appropriate interpretations and decisions made from obtained scores (Messick, 1989; 1995).

Variance. An estimate of the average spread, or variability, or a set of data (Field,

2009).

2. CHAPTER 2: REVIEW OF THE RELEVANT LITERATURE

The present chapter will provide a review of relevant literature regarding the importance of psychometrics when using classical test theory to assess psychological and educational constructs. Different types of psychometrics that can be evaluated through research will be described followed by a review of the most common statistical procedure used to assess those psychometric properties, referred to as simple correlation. Psychometric properties are typically evaluated during the development of a new measure however a brief discussion of the means of assessing the psychometrics of measures already in use will be provided next. A gap in the literature will then be identified where scores from a measure that is in wide use are being interpreted regardless of the absence of psychometric evidence available to support such decisions. The chapter will then close with a brief review of the current study as well as the study hypotheses.

2.1. The Importance of Psychometrics of Measures

Some of the most important tools available to researchers and practitioners are published tests that allow one to evaluate, measure or classify participants according to the underlying construct of interest, also termed the *latent variable* (DeVellis, 2003). Without these measures, researchers would be unable to determine whether or not a treatment had an impact on participants (i.e., a change in scores on the outcome measure following the treatment) or whether the treatment led to no change in the participants (i.e., the same or similar scores on the outcome measure following the treatment). Without valid and reliable interpretations of scores obtained on measures, practitioners and educators may make inaccurate or erroneous decisions as the results obtained on the questionnaire may not accurately measure the latent variable as intended. The potential consequences of making unreliable and/or invalid interpretations of scores

obtained on measures could be damaging, especially in high-stakes situations (Messick, 1995). For example, a practitioner may decide that his or her client is not in need of medicinal intervention for depression based on the client's score on a new depression measure that is lacking reliability and validity evidence. Had the client completed a different measure of depression that had substantive reliability and validity evidence however, the clinician would not only have increased confidence in the scores obtained but also a more accurate interpretation of the client's level of depressive symptoms. The client would then be more likely to receive the appropriate treatment that he or she needed before potentially dire consequence occurred (e.g., depression worsens leading to suicide). Educators also have to ensure that the measures they choose to use produce highly interpretable scores (i.e., strong validity and reliability evidence; Cronbach, 1951) otherwise students are being done a disservice. For example, measures within the education system can be used to classify individuals (e.g., learning disabled), pass or fail individuals, as well as provide information regarding the level of proficiency individuals have across the country. In all three of these examples of test score use, there are a number of invalid decisions and interpretations that could be made if the measure(s) used yielded unreliable scores – learning disabled students may go unassisted, individuals with inadequate performance may be passed, and certain districts may be not be provided with the governmental resources required.

Measuring psychological constructs is not as simple as measuring physical characteristics. Some psychological constructs that researchers and/or practitioners may want to measure cannot be tapped through overt questioning (e.g., questionnaires, proficiency tests). Rather, measurements of behaviour resulting from specific procedures (e.g., interviewing procedures) or manipulations (e.g., the Deese-Roediger-McDermott [DRM] false memory paradigm, Deese, 1959; Roediger & McDermott, 1995; RIF, Anderson et al., 1994) are sought

out. Regardless of the method used, if an observation of behaviour results in a score that is intended to reflect a certain amount of the construct of interest in the individual being observed, then *measurement* has taken place (Crocker & Algina, 1986). With the use of *classical test theory* (or *classical true score theory;* Allen & Yen, 1979; DeVellis, 2003; Kline, 2005; Traub, 1997), researchers and practitioners can evaluate measures to determine their degree of confidence in the reliability and validity of the scores obtained on the measure. The tenets of classical test theory are discussed next.

2.2. Classical True Score Theory

Classical true score (test) theory of measurement began its development when academics acknowledged that human error was likely involved in all human measurement (Traub, 1997). When using classical true score theory of measurement, it is assumed that every measurement is the sum of an individual's *true score, random error* and *systematic error* (Allen & Yen, 1979; Crocker & Algina, 1986; DeVellis, 2003). An individual's *true score* represents the *real* measure of the latent variable that is free from error. *Random error* refers to influences on the obtained test score that are unaccounted for and affect some, but not other, examinees' scores. Random error then impacts individual scores differentially rather than impacting all scores in the same manner as with *systematic error* which systematically raises or lowers the scores of all examinees (and subsequently the mean for the group; Allen & Yen, 1979; DeVellis, 2003). Expressed in equation format, *true score theory* is represented as

$$X = T + e_r + e_s \tag{2.1}$$

where X is the obtained score on a measure or test, T equals the true score, e_r equals random error and e_s equals systematic error.

True score theory is important for psychological and educational assessment because it

allows for an estimation of a true score through the shared variance of the sum of the true score and error scores (DeVellis, 2003; Kline, 2005). Measurement of psychological constructs does not provide a pure measure of individuals' true scores and associated error – they are simply theoretical constructs. What can be obtained, however, are estimates of individuals' true scores and error based on their observed score (sum of both error and true ability) on a test that is believed to be a valid and reliable representation of the underlying construct of interest, or latent variable.

Reliability and validity are two very important components of true score theory (DeVellis, 2003). If the scores obtained on a test that is being developed do not reliably indicate individual differences in the intended latent variable, or the test scores do not produce a valid measure of the entire construct of interest, then an estimate of participants' true scores could never be obtained from that measure as it would only assess error or individual differences on a *different* construct (or a combination of both). Thus, with true score theory it is imperative that examinations are conducted to determine psychometric soundness of measures (Messick, 1989). The following section will review the definitions and types of reliability and validity that are most often used in educational and psychological measurement.

2.2.1. Types of Psychometrics

Two of the most widely discussed psychometric properties of measures are *reliability* and *validity*, which are broad categories of related concepts. Depending on the specific research situation, the term *reliability* refers to the consistency, stability and/or precision of scores obtained on a measure either during one administration or across time intervals (Allen & Yen, 1979; Cronbach, 1951; DeVellis, 2003; Gall et al., 2007; Heffner, 2004). *Validity*, on the other hand, refers to the appropriateness, meaningfulness and application of test scores as well as the

usefulness of judgments made from test scores (Allen & Yen, 1979; Gall et al., 2007; Heffner, 2004; Messick, 1989).

Perhaps the easiest illustration of exactly what reliability and validity refer to is accomplished through picturing a target with a bulls-eye center (DeVellis, 2003). The center, or bulls-eye, of the target represents the construct of interest while throwing a dart at the target represents the administration of the test. The goal of psychological and educational assessment would be to hit the center of the target (the intended construct) a number of times. Such a pattern would indicate reliable (repeatable, stable) and valid (hitting the center) measurement. If the assessment consistently hits the same general area of the target a number of times, but the area was *not* the center, or bulls-eye, the measure would be considered reliable (repeatable, stable) but not valid (*not* hitting the intended construct). If the measure *cannot* hit the center of the target (the intended construct) a number of times (*not* repeatable, *not* stable), the measure is neither valid, nor reliable. To further illustrate what the terms reliability and validity refer to, imagine using a weight scale to measure an individual's height. The same number on the weight scale may be obtained from the same individual a number of times indicating reliability in the measurement, however using the weight scores obtained as a measure of height would be an invalid application of the scores.

How can researchers and practitioners determine whether or not the measure/test they use is in fact measuring the latent variable each time the test is used? By calculating reliability and validity estimates using the assumptions of classical true score theory, one is able to evaluate the extent to which measures are in fact reliable and valid. Each of these psychometric properties will be discussed in turn, first with regards to theory and importance of the psychometric, followed by a description of the most widely used estimates of that psychometric property.

2.2.1.1. Theoretical foundations of reliability. Reliability theory (Allen & Yen, 1979; Lord & Novick, 1968) is based upon the true score theory of measurement discussed earlier. By assuming that psychological and educational tests are measuring a true score along with error, the consistency of scores that are intended to reflect a stable trait (e.g., IQ) can be correlated to help elucidate the reliability of the measure (Cronbach, 1951). If it were assumed that only true scores were obtained through measurement, then correlations between multiple administrations would have to be perfect to be considered reliable. If, on the other hand, the true score was not assumed to be obtained in measurement, measurement would be pointless as only an assessment of error ($r = 0$) would be obtained. With classical test theory, *pure* measures of the reliability of individuals' true scores are never obtained, rather, *estimates* of the reliability of true scores are made based on the shared variance of obtained scores (Allen & Yen, 1979; DeVellis, 2003; Frisbie, 2005) typically using Cronbach's alpha (α; Cronbach, 1951). With classical test theory in place, reliability estimates of .50 indicate that 50% of the obtained score is attributable to the true score and 50% of the score is attributable to error. A reliability estimate of .70, on the other hand, can be considered fairly large (Cronbach, 1951) and indicates that 70% of the obtained score can be attributed to the true score with the remaining 30% of the score due to error, and so on – as the size of the correlation increases so does the percentage of the obtained score that reliably reflects the true score (Frisbie, 2005).

Considering the wide variety of reliability estimates that can be obtained in research and how they can be applied (e.g., stability of test scores, equivalence of test scores, internal consistency of test scores; Cronbach, 1951), the discussion will now turn to the four primary types of reliability evidence: inter-rater/observer reliability, internal consistency reliability, test-retest reliability, and parallel- and alternate-forms reliability.

2.2.1.2. Inter-rater/observer reliability. As the name implies, *inter-rater* or *observer reliability* refers to the degree to which different raters' or observers' scores on the same measure agree, or are replicable (i.e., reliable; Cohen, 1960). To illustrate, consider a research study where children's aggressive behaviour towards dolls is assessed through observation. A specific set of criteria defining exactly what constitutes "aggressive behaviour" is developed and two raters are trained to use the criteria. In order to determine whether or not the two raters are in fact using the criteria in the same manner, and are rating their observations similarly, a reliability estimate is calculated between the amount of agreement or shared variance there is among Rater A's scores and Rater B's scores. When only two raters are used and the ratings have no natural ordering, Cohen's kappa (κ) statistic (Cohen, 1960) should be calculated. Cohen's kappa statistic is a nonparametric analysis with scores that range from 0 – 1 with higher numbers representing a greater degree of agreement (Cohen, 1960). If the scores obtained by different raters have a meaningful order however, and the scores are normally distributed, Pearson's *r,* ranging from -1 to +1, should be calculated when the rating scale is *continuous*. If however, the rating scale used for the observations are simply ordinal, and the scores are *not* normally distributed, Spearman's rho (δ) should be calculated with the correlation coefficient also ranging from -1 to +1.

From the definition of inter-rater/observer reliability, it is obvious that these estimates are only used when there is more than one rater/observer assigning scores and an estimate of degree of agreement between raters is sought. It is also important to note that having more than one rater/observer in research studies where observations are required to assign scores will increase the statistical power and generalizability of the findings if the raters/observers demonstrate adequate agreement. As true score theory ascertains, human error always plays a role in

measurement thus, by having more than one rater/observer, researchers can examine inter-rater reliability estimates to evaluate whether or not the same concepts are being measured by different individuals and whether the assigned scores are replicable (Cohen, 1960).

2.2.1.3. Internal consistency reliability. *Internal consistency reliability* examines whether or not the items within a scale or measure are homogeneous (i.e., measure one trait, such as a test of simple math skills; Cronbach, 1951; DeVellis, 2003). Thus, internal consistency reliability is clearly suited to evaluate tests that are theoretically homogenous rather than heterogeneous (e.g., an IQ test that includes a measure of verbal reasoning, math and reading) although the internal consistency of subscale scores can also be calculated (Cronbach, 1951).

Internal consistency reliability can be established in one testing situation, thus it avoids many of the problems associated with repeated testing found in other reliability estimates such as test-retest reliability (Allen & Yen, 1979). The most common form of internal consistency reliability is the split-halves reliability estimate (Allen & Yen, 1979) and it is most typically reported using Cronbach's (1951) coefficient alpha (α; Cortina, 1993; DeVellis, 2003). If responses on the test are dichotomous however, Kuder-Richardson's KR-20 analysis would be best suited; see Allen and Yen (1979), DeVellis (2003) and Kuder and Richardson (1937) for discussions.

To evaluate internal consistency reliability using the split-halves method, the measure is split into two parts with an attempt to make the two halves *parallel* (Allen & Yen, 1979). In order for the two halves of a test to be *parallel*, the scores obtained on each half must have the same true score and the same error variance (Allen & Yen, 1979). It is important to note that the two tests will very rarely be perfectly parallel (Cronbach, 1951) so some attempt must be made to make the split halves of the measure as parallel as possible, or at least *essentially tau (τ)*

equivalent (Allen & Yen, 1979). Two halves of a measure that are *essentially τ equivalent* will have the same true score for each half of the test, but *unequal* error variances and therefore the obtained scores will differ by an additive constant (Allen & Yen, 1979). There are a number of methods available to help with creating parallel halves of a measure. For example, on a homogeneous test, a fifty-fifty split could occur or odd items may be selected and grouped together with even numbered items in the second half (Crocker & Algina, 1986). Even more specific attempts at making the measures parallel can be made by using the *matched random subsets* method where items are matched based on similar level of difficulty following the administration of the measure to a pilot group (Allen & Yen, 1979; Crocker & Algina, 1986). Regardless of the method, some effort should be devoted towards making the split-halves parallel – if some devoted effort has been made, it is likely that the two halves will at least meet the standards to be considered essentially τ equivalent (Allen & Yen, 1979; Crocker & Algina, 1986).

If the halves of the test are in fact parallel, the internal-consistency reliability estimate of the entire test can be calculated using the *Spearman-Brown formula*. Longer tests typically yield slight overestimates of the true reliability (i.e., upper-bound estimates of reliability) of the measure when compared to shorter tests (Crocker & Algina, 1986; Cronbach, 1951), thus the Spearman-Brown prophesy formula can be applied to correct underestimates of reliability for each half of the test (Crocker & Algina, 1979). If the halves are essentially τ equivalent, the internal-consistency reliability of the entire test can be calculated using coefficient α (Allen & Yen, 1979; Crocker & Algina, 1986; Cronbach, 1951; DeVellis, 2003). In other words, if the variability within each half of the test is equal, then the Spearman-Brown formula can be used to calculate internal consistency reliability of the entire test. If the variance of scores for each half

of the test is somewhat unequal however, Cronbach's coefficient alpha should be used to calculate internal consistency reliability (Allen & Yen, 1979; Cronbach, 1951; DeVellis, 2003; Traub & Rowley, 1991). Essentially, the two halves of the test are alternate forms of each other; thus if coefficient alpha produces a high value (e.g., .70; Cronbach, 1951), then the internal consistency reliability of the entire test/measure is high. If the value is low however, the internal consistency reliability of the entire test is low or poor (Cronbach, 1951).

As with most statistical procedures, there are some limitations to the split halves internal consistency reliability estimate. Sometimes it is difficult, if not impossible, to split a measure into comparable halves. For example, there are instances when items on a test cannot be measured independently, such as with IQ tests that provide total scores based on a number of domains and thereby cannot be separated. Interview protocols or specific procedures, such as that used in the current research, also cannot be separated into comparable halves. In such instances, other forms of reliability estimates, such as alternate-forms reliability or test-retest reliability, should be pursued (Allen & Yen, 1979). These two reliability estimates which are used in the current research are discussed next.

2.2.1.4. Test-retest or stability reliability. Testing the same participants twice with the same test, and then correlating the results provides a *test-retest reliability estimate*, *stability reliability estimate* or *coefficient of stability* (Allen & Yen, 1979; Crocker & Algina, 1986). Pearson's *r* is the most commonly used statistical procedure to evaluate test-retest reliability (Allen & Yen, 1979; Crocker & Algina, 1986; DeVellis, 2003); thus, stability reliability estimates can range from -1 to +1. If all participants obtained exactly the same scores on both administrations of the measure, then perfect reliability would be obtained and there would be a perfect, positive correlation between the scores ($r = 1$; Allen & Yen, 1979; DeVellis, 2003). As

classical true score theory attests, all psychological measurement includes some degree of error (Allen & Yen, 1979; Crocker & Algina, 1986) so perfect correlations between two administrations of the same measure would be extremely unlikely to occur. What test-retest reliability estimates can provide however, are estimates of the ratio of true score variance to observed score variance, or the amount of variance shared across the two administrations of the measure (Allen & Yen, 1979).

When evaluating test-retest reliability of tests and measures, an issue of concern arises with regards to re-administering the same test twice to participants. It is easy to understand how *carry-over*, or *practice,* effects may skew the results. *Carry-over* effects refer to changes in performance due to previous experience with the same test (Allen & Yen, 1979). With previous experience of the same items from the same test, participants may remember and choose the correct (or desired) answer the second time around resulting in an overestimate of the reliability of the scores obtained (Allen & Yen, 1979; Crocker & Algina, 1986; Traub & Rowley, 1991). Thus, test-retest reliability estimates based on two administrations of the same test are often quite high (e.g., .80, Crocker & Algina, 1986). Conversely, participants may be bored and not as interested in the second administration of the test or may be less cooperative the second time around resulting in an underestimate of the reliability of the test (Crocker & Algina, 1986). To help circumvent this issue, a second test may be developed or chosen for administration that is matched to the initial measure in every possible way. By correlating these two measures, a *coefficient of stability and equivalence* (Crocker & Algina, 1986; Cronbach, 1951) is produced that will allow an evaluation of the potential impact of carry-over and practice effects in the coefficient of stability (Crocker & Algina, 1986; Cronbach, 1951).

A second consideration with regards to test-retest reliability involves the length of time

between administrations of the same test (Allen & Yen, 1979, Cook & Beckman, 2006; Crocker & Algina, 1986; DeVellis, 2003). Using a very brief delay between administrations could increase the impact of carry-over effects due to things such as memory, mood or practice while a long interval may reduce the impact of practice but increase the impact of changes in mood or knowledge/information (Allen & Yen, 1979; Crocker & Algina, 1986). To help determine the most appropriate time interval between administrations of the same test to evaluate test-retest reliability, one must consider the underlying trait that the test is intended to measure. For example, if the underlying trait is one that changes over time (e.g., gains due to maturity), then long intervals between testing would likely reduce the reliability due to actual changes in the trait, rather than problems with the measure (e.g., as infants age, new skills develop). In situations where the trait is expected to change over a brief period of time, short testing intervals should be used (e.g., one day to two weeks; Crocker & Algina, 1986). If the underlying trait that one is intending to measure with the test is a theoretically stable one (e.g., proficiency in simple math), then longer testing intervals (e.g., one month to one year) would be appropriate (Allen & Yen, 1979; Crocker & Algina, 1986). Thus, test-retest reliability estimates are most appropriate for measuring the reliability of traits that are not especially susceptible to carry-over effects and are stable across the time interval used.

In situations where the issues associated with administering the same test twice are most prevalent (e.g., in traits that are not stable across time), parallel-forms and alternate forms reliability should be employed. These two methods of evaluating reliability will be discussed next.

2.2.1.5. Parallel- and alternate-forms reliability. Although upon initial consideration, parallel- and alternate-forms reliability seem rather similar to internal consistency reliability, an

important distinction can be made between them. Internal consistency reliability is concerned with whether or not items from a single measure are tapping the same construct. *Parallel-forms reliability*, on the other hand, compares scores on two tests that are strictly *parallel* (i.e., identical alphas, means and variances; Crocker & Algina, 1986; DeVellis, 2003). Thus, adequate parallel-forms reliability estimates would be evidenced through strong, positive correlations between the scores on the two parallel tests (e.g., r =.70). *Alternate forms reliability* estimates however, compare scores obtained on an alternate form of the same measure that purports to tap the same (or similar) underlying construct as the measure being evaluated (Allen & Yen, 1979; DeVellis, 2003). Thus, as with the other reliability estimates described, when establishing good alternate-forms reliability, strong positive correlations are expected (Crocker & Algina, 1986).

Given the definition of parallel forms reliability provided above, there are instances when it may be quite difficult to find a measure that can be used as a parallel form of the measure one is evaluating or developing. To follow the assumptions associated with strictly parallel tests, each item within each of the tests must share an identical relationship with the latent variable (i.e., measures the latent variable to the same degree) and must also contain an identical amount of error (Allen & Yen, 1979; Crocker & Algina, 1986; DeVellis, 2003). Clearly, obtaining two identical tests that meet these assumptions could prove to be very difficult, if not impossible. Thus, researchers and practitioners often turn to *alternate forms* reliability estimates to evaluate the reliability of test scores.

As evidenced through the discussion of the different forms of reliability estimates, each estimate is accompanied with its own considerations and sources of error (DeVellis, 2003). Thus, in order to determine that the scores obtained on a test are in fact reliable, multiple sources of reliability evidence should be evaluated. Each estimate obtained will inform researchers and

practitioners regarding the functioning of the test and latent variable under different circumstances (e.g., different items) within the same individuals. Accounting for as much error in measurement as possible by establishing a comprehensive account of reliability through various forms of reliability estimates will lead to more confidence in the conclusions drawn and decisions made based on those scores. For this reason, whenever possible, obtaining multiple sources of reliability evidence for test scores is preferred.

As discussed earlier, test scores may be reliable without being valid. For example, a test may consistently produce the same score for the same individual across time (e.g., test-retest reliability). However, if the test that the individual is completing is not actually measuring the latent variable that it purports to measure, the test is rendered invalid – the inferences, decisions and consequences resulting from the test scores would be based on erroneous data. The specific issues involved with evaluating the validity of tests scores are discussed next.

2.2.1.6. Theoretical foundations of validity. The second important psychometric property of measures is validity. Messick (1989; 1995) describes *validity* as "an overall evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores or other modes of assessment" (p. 5). Thus, principles of validity, like that of reliability, apply to all types of assessments that assign scores based on observed consistencies in behaviour (Messick, 1989, 1995). Various types or forms of validity evidence have been identified in the literature that fall under three broad categories: construct validity, criterion-related validity and content validity. The most common methods for evaluating these three categories of validity will be discussed next.

2.2.1.7. Construct validity. *Construct validity* can be defined as the extent to which a

measure actually measures what it intends to (Allen & Yen, 1979; DeVellis, 2003; Gall et al., 2007; Heffner, 2004; Messick, 1989). For example, a valid measure of job performance may be the amount of sales made by an individual within a month. If an employer views sales performance as an important aspect of job performance then measuring that characteristic and using the obtained scores to make a decision would constitute a valid application of the scores. An invalid measure of job performance however would be shoe size. There is no expected relationship between shoe size and job performance thus using shoe size to predict future job performance would be an invalid application of the observed scores. Obtaining a measurement of shoe size may be highly reliable (replicable), however, obtaining a score reliably does not automatically imply that the score is being validly used.

Convergent and discriminant validity are both aspects of construct validity (Messick, 1989; 1995). *Convergent validity* refers to the extent to which scores on a measure share a high, medium or low relationship with scores obtained on a different measure intended to assess the same or a similar construct (Messick, 1989; 1995). For example, if a new measure of self-esteem was being developed, participants could complete an already psychometrically sound measure of self-esteem as well as the new measure. Correlating the scores obtained on both measures would provide convergent validity evidence allowing for an interpretation of the overall validity of the scores obtained on the new measure (Messick, 1989; 1995). *Discriminant validity* on the other hand, refers to the extent to which scores on a measure *do not* share a relationship with scores obtained on a theoretically unrelated measure (Messick, 1995). Evidence of discriminant validity could be obtained in the above mentioned example by having participants complete a measure of social desirability in addition to the new self-esteem measure. Correlations between social desirability scores and self-esteem scores should be weak if the

measures are providing assessments of two different constructs.

Evaluating construct validity is more of an ongoing process than a single statistic (Allen & Yen, 1979; Cook & Beckman, 2006; DeVellis, 2003; Messick, 1989) that develops as new theoretical and practical advancements are made regarding the construct being measured. To establish construct validity, current theory and empirical findings regarding the construct of interest must be surveyed. Predictions based on theory and empirical findings can be tested through research with the measure and, depending on the results, it could be concluded that the measure either does or does not demonstrate acceptable validity. The primary method of evaluating construct validity is through repeated testing and empirical research. If researchers, who are considered experts in their field, continue to use a measure and are able to meaningfully discuss and interpret the results in relation to theory and predictions, then that in itself is evidence of construct validity (Messick, 1989). There is always a chance however, that human error (and subsequently measurement error) has come into play and research findings that may seem to be evidence of construct validity are actually due to error. Conversely, evidence that appears to demonstrate that a measure has poor construct validity may also be due to error. As with reliability then, multiple forms of validity evidence should be obtained in order to ensure confidence in, and the appropriate use of, test scores.

2.2.1.8. Criterion-related validity. The second category of validity is criterion-related validity. *Criterion-related validity* can be used when scores obtained on a test can be related to some criterion (Allen & Yen, 1979; Messick, 1989, 1995). For example, if students with high scores on the Graduate Record Examination (GRE) actually achieved higher grade-point averages in graduate school when compared to students who obtained low GRE scores, the GRE would be said to have good criterion-related validity. In other words, individuals' scores on the

GRE would be able to *predict* their future graduate school performance.

Criterion-related validity estimates can be obtained in one of two ways, depending on the time interval between the administration of the measure (e.g., GRE) and collection of the criterion measure (e.g., graduate school grade-point average; Allen & Yen, 1979). The goal of *predictive validity* is to predict future behaviour from scores obtained on a measure (Allen & Yen, 1979). A predictive validity estimate would be established in the GRE example by first administering the GRE to a group of graduate students entering their first year of graduate studies. After a reasonable amount of time has passed to allow for a reliable examination of their performance in graduate school, a measure of the criterion (e.g., academic average of students' first year of graduate school) would then be collected (Allen & Yen, 1979; Cook & Beckman, 2006; Crocker & Algina, 1986; Messick, 1989). Students' GRE scores (the predictor) and their criterion scores would be correlated resulting in a *predictive validity coefficient*. Measurement of the criterion is sometimes collected at, or around, the same time as collection of the predictor however, and in these instances a *concurrent validity coefficient* is calculated. For example, consider a researcher who has developed what she believes to be a briefer predictor of driving performance and who now wants to collect evidence of its criterion-related validity to compare to the current, less efficient, method. To accomplish this, she could administer her new measure (predictor) to a group of driver education students along with the standard measure and then collect the criterion measure of driving performance on the same day. Correlating these scores would provide concurrent validity evidence and would indicate whether adopting the new, more efficient predictor is warranted. As with typical correlation coefficients, a higher correlation coefficient (e.g., $r = .70$) demonstrates strong criterion-related validity while correlations at or near 0 indicate no criterion-related validity.

2.2.1.9. Content validity. Content validity is most often evaluated during the development of a measure rather than after the measure has been created. *Content validity* refers to the degree to which the measure comprehensively assesses the underlying construct of interest (Allen & Yen, 1979; DeVellis, 2003; Messick, 1989; 1995) and can be grouped into two types: *face validity* and *logical validity* (Allen & Yen, 1979). *Face validity* refers to the degree to which the content of a measure appears to comprehensively assess all the relevant domains of the latent variable and *logical validity* is a more sophisticated version of face validity (Allen & Yen, 1979). There are no set statistical techniques for evaluating content validity as there are for criterion-related or construct validity (Allen & Yen, 1979; Messick, 1986); however, there are guidelines to aid in the development of a measure that could be argued to have strong content validity. For example, three to five judges who are experts in the field of interest are asked to evaluate whether or not the construct being tapped is fully represented through all the items on the measure or if some items are missing or are redundant (DeVellis, 2003). Expert judges may also be asked to rate the 'fit' of each item on the measure to its respective domain using a Likert-type rating scale. Examining the variance in judges' fit ratings would then allow the measure developer to assess the degree of agreement for item fits across all domains and judges. Another option to help ensure that a newly designed measure has strong content validity is to clearly define the different domains that are believed to compose the construct(s) being assessed by the measure (Allen & Yen, 1979). A number of items that appear to tap each of the defined domains could then be written further supporting the content validity of the measure (Allen & Yen, 1979).

2.2.2. Summary

Determining whether or not a measure yields reliable and valid scores should be of primary concern for users. Reliability and validity evidence are not only important when

developing new measures but are also important when selecting measures to use in research or to use in educational and psychological settings. Imagine that a researcher received a grant to fund her research but did not choose a valid and reliable measure of an important outcome variable. Her research regarding this variable would be rendered almost useless as any applications, interpretations or conclusions made based upon the unreliable scores would be erroneous – she could not conclude that a measurement of participants' true scores for that variable occurred. Or, imagine that a final job applicant was accepted for hire based on an unreliable or invalid score obtained on a job performance measure – the manager would have lost a valuable employee and related sales due to choosing a poor measure of job performance. Regardless of the situation, maximum accuracy in measurement is always the goal, thus choosing the most valid and reliable measure of the construct of interest helps to ensure just that - maximum accuracy. Most analyses that produce reliability and validity estimates of measures employ the use of simple correlation, or Pearson's *r*. For this reason, a brief review of simple correlation is provided next.

2.3. Using Simple Correlation in Research

Examining the strength of association or the relationship between two variables, or *correlation* (Field, 2009), is quite a common statistical procedure. For example, a researcher may want to investigate whether or not individuals' depression scores share a relationship with their self-esteem scores. By administering both a depression measure and a self-esteem measure to the same participants and correlating the test scores (e.g., using Pearson's *r*), the researcher is able to examine the potential relationships between these two variables (Field, 2009). When conducting correlations between two variables, the correlation coefficient produced (ranging from +1 to -1) can show that the variables are related in one of three ways: (a) a *positive relationship* may exist where an increase in variable A leads to an increase in variable B, (b) a

negative relationship may exist where an increase in variable A leads to a decrease in variable B, or (c) *no relationship* between the variables may exist meaning that scores obtained on the two measures are not related (Field, 2009). Continuing with the previous example, a positive relationship between depression scores and self-esteem scores would be illustrated through a strong, significant positive correlation (e.g., *r*'s = .70 to .90; Hinkle, Wiersma & Jurs, 2003) – as individuals' depression scores increased, so did their self-esteem scores. A negative relationship however, would be evidenced through a strong, significant negative correlation (e.g., *r*'s = -.70 to -.90; Hinkle et al., 2003) – as individuals' depression scores increased, their self-esteem scores decreased (or vice versa). No relationship between the variables however would be demonstrated through a non-significant correlation around $r = 0$ (Field, 2009). It is important to note that statistically significant correlations do not indicate causation (Field, 2009). What is being examined through correlation is the existence or non-existence of a *relationship* between variables – if a relationship exists, one cannot claim that changes in one variable caused changes in the other. What can be claimed however, is that some sort of relationship (e.g., linear) exists between the variables (Field, 2009).

In order to understand correlation, a discussion of covariance is needed. *Covariance* refers to the extent to which change in one variable is met with change in the other variable (Field, 2009). In other words, in order to examine whether or not two variables share a relationship, one must determine whether or not an increase in variable A leads to an increase (or decrease) in variable B – the degree to which the scores *covary* is of interest (Field, 2009). In order to evaluate these potential relationships, an examination of the *variance* of both sets of scores is required (Field, 2009). That is, if a relationship exists between the scores obtained on two measures, when a score on one measure deviates from the mean of all scores for that

measure, a similar (or inverse) deviation from the mean for the other measure would also be expected (Field, 2009). Establishing covariance can be a useful means of assessing the potential relationships between two variables, however the covariance obtained is completely dependent upon the scales of measurement used – in other words, covariance is not a standardized measure (Field, 2009). This is problematic due to the inability to compare the strengths of relationships across different scales (Field, 2009). In order to overcome this issue, covariance is converted into a standard set of units referred to as *standardization* (Field, 2009) by dividing the covariance by the product of the standard deviations obtained for both measures (Field, 2009). Expressed in equation format, correlation can be calculated through:

$$r = \frac{\text{covary}_{xy}}{s_x s_y} \quad \text{or} \quad r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(N-1)\, s_x - s_y} \tag{2.2}$$

where x_i is the data point for variable A,
$\bar{x}$ is the sample mean for variable A,
y_i is the data point for variable B,
$\bar{y}$ is the sample mean for variable B,
N is the number of observations,
s_x is the standard deviation for variable A and
s_y is the standard deviation for variable B.

The resulting statistic is referred to as *Pearson's correlation coefficient* (Field, 2009) and is also often used as an indication of effect size. *Effect size* refers to a standardized measure of the magnitude of an observed effect (Field, 2009). There are no firmly set standards for interpreting the size of an effect obtained but some 'rules of thumb' are available to aid interpretation. For example, one statistical textbook (Field, 2009) recommends interpreting *r*'s ±.10 as small effects, *r*'s ±.30 as medium effects, and *r*'s ±.50 as large effects. A similar statistics textbook recommends interpreting the size of a correlation based on the following more defined ranges: ± .90 to 1 should be interpreted as a very high correlation, .70 to .90 as a high

correlation, .50 to .70 as a moderate correlation, .30 to .50 as a low correlation and 0 to .30 as a little, if any, correlation (Hinkle et al., 2003).

As mentioned earlier, correlation does not indicate causation, thus a second useful statistic that can be obtained during correlational analysis is the coefficient of determination. Squaring Pearson's *r* yields the *coefficient of determination* which describes the amount of variance in scores on one measure that is shared by the other (e.g., $r = .24$, $r^2 = .058$, thus the two measures share 5.8% of the variance; Crocker & Algina, 1986; Field, 2009). Using the coefficient of determination allows researchers to examine the magnitude of the relationship between variables thereby facilitating an interpretation of the practical significance of the results (Field, 2009) rather than limiting themselves to statistical significance only (Cicchetti, 2001; Kirk, 1996).

Considering the wide array of different types and forms of reliability and validity evidence and the different information provided through each estimate, the best practice is to examine as much reliability and validity evidence as is available (Crocker & Algina, 1989; Messick, 1989). Typically, some validity and reliability evidence is established during the development of a new measure, however there are instances when little or no evidence is available in the literature regarding its use. Simple correlation would still likely be used to examine the psychometrics of the measures in such instances, but only if the need for such research is realized. An example of a study that has examined the psychometrics of a measure after it has already been in use is provided next followed by a review of the literature surrounding the primary measure that will be used in the current research, RIF (Anderson et al., 1994). The present chapter will then close with a brief review of the current study along with the study hypotheses.

2.4. Establishing Psychometrics of Measures Already in Use

Reliability and validity evidence often accumulates over time as theory develops and more researchers employ the measure in their studies, but there are also instances when a certain pattern of results emerges from a manipulation that appears robust, leading researchers to interpret the scores as a measure of a stable individual difference in some underlying construct. One such example of this examination of the psychometric properties of a manipulation can be found in the recent publication by Blair, Lenton and Hastie (2002) examining the reliability of the Deese-Roediger-McDermott (DRM for Deese, 1959; Roediger & McDermott, 1995) false memory paradigm as a measure of individual differences in susceptibility to false memories. The occurrence of false memories in the context of associated word lists was first introduced by Deese in 1959 and was revived in the empirical literature by Roediger and McDermott in 1995. In the DRM paradigm, participants are presented with lists of semantically associated words to study (e.g., hot, snow, warm, winter, ice) that all converge on a common unpresented associate (e.g., cold). During recall, participants often report the unpresented item (e.g., cold) and report confidence in their decision that it had been studied thereby demonstrating a *false memory* for that item. Research examining the conditions and situations that may moderate or impact the false memory effects obtained through the DRM paradigm began appearing in the literature (e.g., McDermott & Roediger, 1998) as did research regarding the characteristics of false memories generated through the procedure (e.g., Mather, Henkel & Johnson, 1997). Without empirical evidence supporting that individuals' susceptibility to DRM false memory was a stable individual difference, researchers began examining different groups of people who may theoretically be more or less susceptible to DRM false memories than others (e.g., individuals with Alzheimer's, Balota, et al., 1999). Research using the DRM paradigm was initiated in 1959

(Deese, 1959), revived in 1995 (Roediger & McDermott) and continues today, but it was only in 2002 when Blair et al. provided the research literature with empirical evidence demonstrating that the false memory scores produced through the procedure were in fact quite stable across testing intervals. Had Blair et al. demonstrated that the false memory scores produced through the paradigm were not stable across time or individuals then all DRM studies where the false memory results were interpreted through an individual difference perspective would have to be re-evaluated. For example, although individuals with Alzheimer's dementia have been shown to be more susceptible to DRM false memories (Balota et al., 1999), if the false memories produced through the paradigm did not reflect a reliable individual difference in false memory susceptibility, different scores (e.g., less susceptibility) would likely be obtained from the same individuals on different testing days. Subsequently, the interpretations, decisions and/or conclusions based on the unreliable scores would be in error as well.

Clearly, establishing the reliability and validity of psychological and educational assessment measures should be considered an important 'first step' in their development, however as illustrated in the DRM false memory example above, this is not always the case. A similar instance where there is a dearth of research examining the psychometrics of a measure that is prolifically evident in the literature is that of *retrieval-induced forgetting* (RIF; Anderson et al., 1994). During the typical RIF procedure (e.g., Anderson et al., 1994), participants first study a number of categories of information (e.g., two categories of sentences or words) that are all semantically related to the assigned category cue word. During this study phase, each item that is to be remembered is presented one at a time, in random order alongside the category cue (e.g., fruit – orange, vegetable – carrot, fruit – apple, vegetable – beans, fruit – strawberry, vegetable – onion, fruit – pear, vegetable - peas). Once all items have been presented,

participants engage in a retrieval-practice phase which involves retrieving *some* of the items from half of the categories three times each using a fill-in-the-blank task (e.g., fruit – or____, fruit – ap____). Following a brief distractor task, participants are presented with the category cues, one at a time in random order, and are asked to recall all of the items they remember studying that were paired with that cue during the study phase.

Participants' final recall data following the classic RIF procedure demonstrates typical practice effects for Rp+ details – items that received extra practice are recalled significantly more often than the unpracticed, baseline details. What is of specific interest however, is the *inhibition* or *forgetting* of the unpracticed details (Rp-) that are semantically associated to the practiced details (Rp+). The act of repeatedly retrieving some of the details during the retrieval-practice phase is argued to lead to mental competition between the practiced and unpracticed items from the same semantic category (Anderson, 2003; Anderson et al., 1994) resulting in inhibition of the unneeded (i.e., unretrieved), related details (Rp-). Thus, when examining the three levels of practice used during the RIF procedure (Rp+, Rp-, NRp), Rp+ details are recalled significantly more often than both NRp and Rp- details with significantly lower recall of Rp- details when compared to the NRp baseline condition that received no practice manipulation.

Over 15 years of research have been devoted to the RIF paradigm since its initial appearance in the literature. Typical RIF effects (i.e., significantly higher recall of Rp+ details and significantly lower recall of Rp- details when compared to the NRp baseline details) have been extended beyond simple word lists (Anderson et al., 1994) to sentences (e.g., Gómez-Ariza, Lechuga, Pelegrina, & Bejo, 2005), social stereotypes (Dunn & Spellman, 2003), social metacognitive judgments (Storm, Bjork, & Bjork, 2005), facts about the self (Marche et al., 2011), and others (Macrae & MacLeod, 1999), as well as autobiographical (Barnier, Hung &

Conway, 2004) and eyewitness (MacLeod, 2002; Migueles & García-Bajos, 2007; Shaw et al., 1995) memory, to list only a few examples. RIF has also been found in child (Ford et al., 2002; Marche et al., 2011), adult (Barnier et al., 2004; Aslan et al., 2007) and older adult (Aslan et al., 2007) populations, is being examined as a potential memory-based intervention (Marche et al., 2011; Wessel & Hauer, 2006) and as an individual difference in normal (Johansson, Aslan, Bäuml, Gäbel & Mechlinger, 2007) and clinical populations (e.g., Harris, Sharman, Barnier & Moulds, 2010; Moulin, Perfect, Conway, North, Jones & James, 2002).

Interest in the forgetting produced through the RIF procedure persists to the present day and researchers continue to develop theory regarding RIF in adult populations (e.g., Anderson, 2003; Anderson et al., 1994; MacLeod, Saunders & Chalmers, 2010) and to identify certain boundary conditions of the effect (Anderson & McColluch, 1999; Bäuml & Kuhbander, 2007). It has been argued that cognitive inhibition is the underlying process driving RIF (e.g., Anderson, 2003; Anderson et al., 1994; Barnier et al., 2001; MacLeod, 2002). *Cognitive inhibition* refers to individuals' ability to suppress previously activated cognitive processes or representations (Aslan & Bäuml, 2011; MacLeod, 2002; Wilson & Kipp, 1998). It is thought to develop during childhood at age 7 or 8 (Harnishfeger & Bjorklund, 1994; Harnishfeger & Pope, 1996; Wilson & Kipp, 1998), differ across individuals (Harnishfeger & Bjorklund, 1994) and facilitate performance on many cognitive tasks (Harnishfeger & Pope, 1996). However, contrary to the inhibition account of RIF, children as young as 7 years of age have demonstrated typical RIF effects (e.g., Ford et al., 2004; Marche et al., 2011). Although there is a continuing debate regarding the most appropriate account of the underlying processes driving RIF, research examining the psychometric properties of the forgetting induced through the procedure would further theory and lead to accurate and appropriate interpretations of the scores obtained. Indeed,

much has been learned regarding the fundamentals of RIF (e.g., Anderson, 2003; Anderson et al., 1994; MacLeod, Saunders & Chalmers, 2010), however as of yet, it is unknown whether or not the degree of forgetting induced through the RIF procedure is a reliable and valid individual difference measure of forgetting ability.

2.5. The Current Study

To illustrate how reliability and validity evidence can be empirically evaluated, and to address a clear gap in the literature, the purpose of the current thesis is to examine a number of psychometric properties of RIF scores. Two sets of 60 category – word pairs (e.g., Fruit – banana, Fruit – apple) were equated according to semantic relatedness to the items' respective categories and one set of facts about two fictitious islands were obtained from past research (e.g., Bilu – The main crop on Bilu is corn, Bilu – Most houses on Bilu are made of wood; Macrae & MacLeod, 1999). Across three phases, participants were asked to complete all RIF tasks once, as well as to repeat one of the category – word pair RIF tasks (Phase II) and the facts RIF task (Phase III). *Test-retest reliability* of the degree of forgetting obtained through the RIF procedure was evaluated by correlating both the matched and repeated category – word pair RIF scores with one another while *alternate forms reliability* evidence was evaluated by correlating all category – word pair RIF task scores with facts RIF task scores. Obtaining more psychometric evidence is better than less so both the convergent and discriminant validity evidence of forgetting scores obtained through the RIF procedure were also evaluated. *Convergent validity* evidence was examined by having participants complete a self-report measure of everyday cognitive failures and correlating these scores with RIF scores and *discriminant validity* was evaluated though correlations between participants' RIF scores and scores on a social desirability measure.

2.5.1. Hypotheses

1. Typical RIF effects are expected to be obtained for all RIF tasks. Given the robustness of the RIF effect (e.g., Anderson et al., 1994; MacLeod, 2002), typical RIF effects are expected for all RIF tasks. That is, in addition to practice effects for the Rp+ items (i.e., significantly higher recall of Rp+ pairings compared to both Rp- and NRp), participants are expected to recall significantly less Rp- pairings compared to the NRp baseline.

2. RIF scores (the degree of forgetting demonstrated) are expected to be positively correlated across all RIF tasks indicating stability reliability. Current theory regarding RIF postulates that *cognitive inhibition* is responsible for the forgetting induced through the procedure (e.g., Anderson, 2003; Anderson et al., 1994). Researchers argue that the repeated retrieval of selected details creates competition between other details from the same semantic category (Anderson et al., 1994); it is this competition at retrieval that leads to the temporary forgetting induced through the procedure (Anderson et al., 1994; MacLeod & Macrae, 2001). Although cognitive inhibition is sensitive to current cognitive demands (MacLeod & Macrae, 2001), individuals' level of inhibition should be relatively stable across short periods of time (e.g., a few minutes to a few weeks). Therefore, if inhibition is in fact a driving mechanism of RIF then significant positive correlations of forgetting scores are expected across all RIF tasks, providing stability reliability evidence.

3. Strong positive correlations are expected between the RIF tasks that participants complete twice, indicating test-retest reliability of RIF scores. Correlations between RIF scores for the tasks that participants completed twice are expected to account for a greater amount of variance due to the impact of carry-over or practice effects. Positive correlations

between RIF scores for the two equated sets of word list RIF tasks are also expected and will provide a test-retest reliability estimate with carry-over effects removed.

4. Positive correlations are expected between the category – word pair RIF tasks and the facts RIF task that participants complete, indicating alternate forms reliability of RIF scores. Correlations between forgetting scores obtained for the word list RIF tasks and the facts RIF tasks will yield estimates of alternate forms reliability. Within the RIF literature, the paradigm has been applied to various types (e.g., eyewitness, autobiographical) and forms (e.g., sentences, words) of memories thereby yielding alternate forms of the RIF effect. Researchers, who are guided by theory, argue that the underlying processes that drive RIF are common, regardless of the materials used (e.g., Anderson, 2003). Thus, if research and theory are correct, forgetting scores obtained using the word list RIF tasks should be highly positively correlated with the alternate forms of the procedure using facts.

5. A significant inverse correlation is expected between RIF scores and scores on a measure of cognitive failures, indicating convergent validity evidence of RIF scores. Correlating scores obtained on a measure of everyday forgetfulness (the Cognitive Failures Questionnaire, CFQ; Broadbent, Cooper, FitzGerald & Parkes, 1982) with all RIF scores will produce a convergent validity estimate. For this correlation, a significant inverse relationship is expected as past research has shown that CFQ scores are inversely related to RIF scores (Groome & Grant, 2010).

6. No significant correlation is expected between social desirability scores and all RIF scores indicating discriminant validity between the measures. Discriminant validity evidence will also be evaluated by correlating RIF scores with scores obtained on a measure that taps an unrelated latent variable - social desirability (Social Desirability Questionnaire – 17, SDS-17;

Stöber, 2001). SDS – 17 scores should not share a relationship with the degree of forgetting participants experience through the RIF procedure, as forgetting and social desirability are considered to be two theoretically separate constructs. Therefore, no relationship is expected to be found when SDS – 17 scores and forgetting scores are correlated (e.g., *r*'s around 0).

3. CHAPTER 3: METHOD

The present chapter describes the specific methodology used to evaluate the study hypotheses and includes a description of the participants who took part in the study, the materials used, the design employed and the execution of the study procedure.

3.1. Participants

A total of 68 participants completed Phase I of the study with 65 of those participants also completing Phase II. Eleven participants erroneously studied the category-word pairs in a non-random order therefore, these participants were excluded from analysis. Three participants withdrew from the study following the first testing session, two participants were older than 45 years and were excluded from analysis to narrow the age range of participants, and one participant was required to take medication during data collection that greatly impacts memory. These six participants were also removed from analysis. The final sample for Phase I and Phase II consisted of 50 participants (M_{age} = 23.96 years, *SD* = 5.17, *range:* 18 – 36 years) with 29 participants being female. Participants who completed all three phases (n = 17; 10 females) had an average age of 26.29 years (*SD* = 4.21). The majority of participants spoke English as their first language (n = 42), two participants spoke Cantonese as their first language and one participant each who spoke Fuijanese, Bengali, Polish, Romanian, Arabic, and Chinese as their first language.

3.2. Design

The current research provides test-retest and alternate forms reliability estimates as well as convergent and discriminant validity evidence. To accomplish this, all participants completed three word list RIF tasks and one facts RIF task counterbalanced across the first two phases of testing. During Phase III, participants completed the facts RIF task again to provide a test-retest

reliability estimate for the alternate forms (facts) RIF task. Each set of word lists was administered twice, two weeks apart, to half of the participants (i.e., half completed word list A twice, half completed word list B twice). The first facts RIF task was administered to half of the participants at Phase I, with the remaining half of participants completing the first facts RIF task at Phase II. Participants were invited back via email to repeat the facts RIF task at Phase III. Refer to Figure 3.1 for a visual depiction of the study design; please refer to Appendix A for a diagram of the full counterbalancing orders.

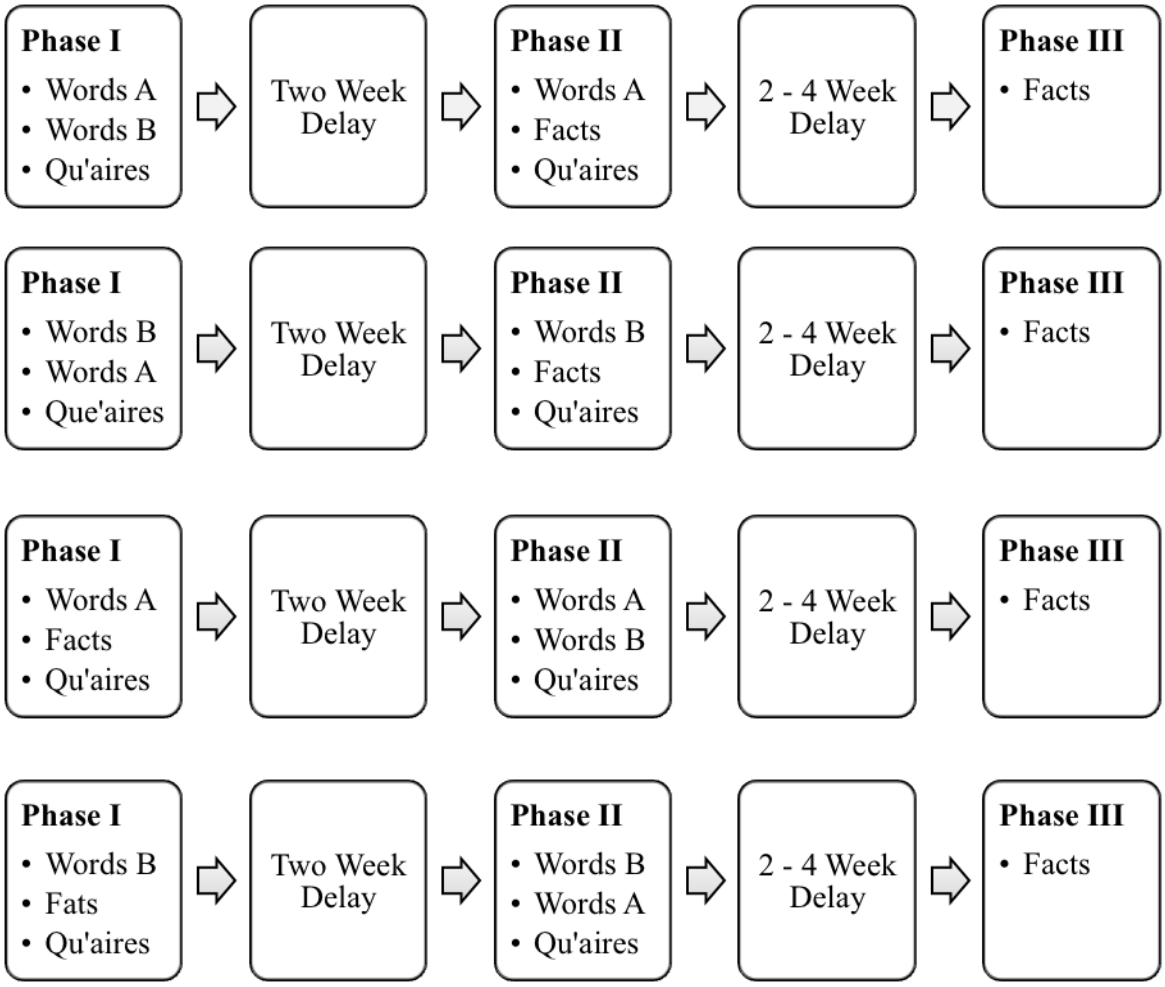


Figure 3.1
Visual depiction of the study design. Each row represents a participant's task order. One set of word lists (A or B) is completed at both Phase I and Phase II. The facts RIF task occurs at either Phase I or Phase II, and is repeated at Phase III. Each questionnaire (Qu'aires) for individual differences is randomly assigned as a filler activity between retrieval-practice and test. Two filler activities involve a visual search task rather than an individual difference measure. The order of tasks within each cell above is counterbalanced across participants (a total of 16 possible orders).

3.3. Materials

As described in the introduction, different psychometric properties require different methodologies and are accompanied by different theoretical considerations. For these reasons, the materials required to establish each estimate are described separately.

3.3.1. Word list test-retest reliability materials. When following *classical test theory* (e.g., Allen & Yen, 1979; Crocker & Algina, 1986) efforts must be made to help reduce or eliminate the potential impact of carry-over and/or practice effects when establishing test-retest reliability (DeVellis, 2004, Allen & Yen, 1979). Thus, two sets of comparable word list materials were required and were created using Anderson et al.'s (1994) pioneer RIF research as a guide. The specific steps taken to develop the current set of word list materials are discussed next.

3.3.1.1. *Item selection.* Prior to selecting the specific categories and words required to create two matched sets of word lists, a number of rules to guide item selection had to first be in place. Three important considerations regarding the development of RIF materials include: the degree of semantic association between category-exemplar pairs, the uniqueness of the letter strings required for retrieval-practice, and cross-category contamination from using exemplars that may fit with more than one category (Anderson et al., 1994). Regarding the first consideration, *category - exemplar association*, past RIF research has demonstrated that the degree of association between categories and corresponding exemplars has an impact on the degree of inhibition obtained through the procedure (Anderson et al., 1994). Typically, the mean output position of items, referred to as the items' *rank,* is used as a measure of the degree of category membership, and greater RIF effects (i.e., more forgetting) have been found for items that share a strong association (i.e., lower ranks) to their category cue when compared to those with weak associations (i.e., higher ranks; Anderson et al., 1994). To obtain these ranks, participants are presented with a category (e.g., Fruit) and are asked to write down all of the words that come to mind after reading the category in the order that they come to mind (e.g., apple, banana, orange; van Overschelde, Rawson & Dunlosky, 2004). The average output

position per item across all participants is then calculated with the resulting score indicating that item's rank in relation to its category (van Overschelde et al., 2004). Thus, to ensure the likelihood of replicating the RIF effect, and to create two sets of comparable word lists to examine test-retest reliability, items within each category were selected based on the items' ranks and *total* proportion of recall (i.e., the proportion of participants who listed the item as a member of the category regardless of output position; van Overschelde et al., 2004). Van Overschelde et al.'s (2004) norms have been validated by Bauer and Gourgouvelis (2009) as comparable to Battig and Montague's (1969) norms that were used by Anderson et al. (1994) in their initial RIF publication.

Prior to selecting and evaluating the ranks and totals of the chosen items, the two other considerations mentioned earlier must be in place. Ensuring that the first few letters of each item in a set of word list materials is unique eliminates the opportunity to contaminate participants' retrieval-practice trials by potentially prompting multiple responses. For example, if the category "Bird" included both "Bluejay" and "Bluebird" as exemplars, participants may believe that they are retrieving the appropriate item during their retrieval-practice trials, when the researcher intended the opposite "Bl____" item to be retrieved. The final consideration, ensuring that items in a set of word lists clearly fit within only one category, is also required to reduce the likelihood of contaminating participants' responses. If, for example, the categories "Fruit" and "Vegetable" were used in the same study list, and the exemplar "Tomato" was listed under one of those categories, participants' final recall data may be contaminated due to the potential for "Tomato" to be categorized implicitly by individual participants as either a "Fruit" or "Vegetable." Thus, during item selection for each category, specific efforts were made to ensure that the first two letters of all items within a set of word lists remained unique and that

each item selected fit within only a single category.

Keeping these considerations in mind, two matched lists consisting of a total of 60 words from 10 categories (six items per category) were selected from recent category norms (van Overschelde et al., 2004). To ensure that the two constructed study lists (see Appendix B for the word lists and retrieval-practice fragments) were comparable to each other with regards to the items' ranks (range 1.4 – 5.8) and total proportions (range .06 - .93), two one-way ANOVAs were conducted which revealed no significant differences according to rank, F (1, 119) = .006, p = .94, or total proportions, F (1, 119) = .881, p = .35. Additional efforts were then made to address the potential issue of cross-category contamination by asking a panel of five judges to evaluate the items' category membership. Each list was randomized and entered into a grid with the different category names beside each item. Judges were then asked to read each item in each of the two lists and circle all of the categories that each item fit into (see Appendix C for the rating form). As anticipated, all items within each list were only assigned to a single category by all judges.

3.3.1.2. *Retrieval-induced forgetting task materials.* All RIF tasks followed the same general template of (a) *study* (study each pair for 5 seconds each), (b) *retrieval-practice* (practice retrieving some of the pairings from memory through a fill-in-the-blanks task, (c) *filler activity* (work on completing questionnaires and/or the visual search task for 5 minutes) and (d) *final test* (cued recall of as many pairings as possible). The materials required to complete these tasks are discussed next.

3.3.1.2.1. *Study booklets*. Each of the two finalized sets of 60 category – exemplar pairs (6 exemplars for each of the 10 categories) was then used to create eight study orders per set. The order of each study book was random with the following constraints applied: (a) one

category – exemplar pair from each of the 10 categories must be cycled through prior to using a pair with the same category again, and (b) the same category must not be studied back to back.

3.3.1.2.2. *Retrieval-practice booklets*. Before creating the retrieval-practice booklets, the category – exemplar pairs that would be used had to be selected. The categories that were to undergo retrieval-practice were selected first by assigning each category a number from 1 – 10. Next, four unique sets of four numbers ranging from 1 – 10 were generated using an online research randomizer (www.randomizer.org) and the category that matched each of the numbers generated was targeted. To select the items within each of the four targeted categories that would undergo retrieval-practice, each item within a category was assigned a number from 1 – 6. Four sets of three digits ranging from 1 – 6 were then generated using the same randomizer and items with a matching number were selected.

These four sets of selected retrieval-practice pairs from each word list were then used to create four retrieval-practice booklet orders per word list set. Each booklet consisted of three trials of nine category – fragment pairs (e.g., Fruit – Or____). The order of fragments within each trial was random with the constraint that the same category could not be used consecutively. Once category – fragment pair appeared on each page by itself.

3.3.1.2.3. *Recall booklets*. Four recall booklets were created for each word list set. One category cue (e.g., Fruit) headed each page of the recall booklets and the order of cues was random. To create the four orders, each category cue was assigned a number from 1 – 10; four unique sets of 10 digits ranging from 1 – 10 were then generated (www.randomizer.org). Each of the four orders was then sorted according to the assigned random digits.

3.3.2. Alternate forms reliability materials. Past research by Macrae and MacLeod (1999) demonstrated RIF effects using facts (10 each) regarding two fictitious islands. A copy of

these materials was obtained from Dr. MacLeod and was used in the current research to obtain both alternate forms, and test-retest reliability estimates (see Appendix D for the facts lists and retrieval-practice fragments). The facts were initially designed with the intent of mirroring the type of information that would typically be found on a geography exam. Fictitious islands were used in order to reduce the chance that general geography knowledge would impact results on the test (Macrae & MacLeod, 1999). The only adjustment made to these previously used materials was to change one retrieval-practice fragment from each list. Macrae and MacLeod originally used one numerical digit as a retrieval-practice fragment on each list (e.g., "9__ of people on Tok own a bicycle) which was changed in the current research to a written word instead (e.g., "93% of people on Tok own a _____").

Material construction for the facts RIF tasks was quite similar to construction of the word list RIF task with a few exceptions. Only two categories, the islands of Bilu and Tok, were used and no equated set of materials was developed. Four different orders each of the study booklets and retrieval-practice booklets were created with the same randomization procedure used for the word list RIF tasks (refer to Appendix D for the facts retrieval-practice fragments). Only two recall orders were created as only two category cues were used. Each page of the two page recall booklet had one of the two island names printed at the top.

3.3.3. Convergent validity materials. The most widely accepted explanation for the occurrence of RIF is that of cognitive inhibition which is argued to be an adaptive trait with the function of overcoming interference (Anderson, 2003). For example, correctly recalling a friend's new telephone number would be extremely difficult if the stored memory trace for the friend's old number was equally as strong as the trace for the new number. In this instance, cognitive inhibition would be argued to prevent access to the friend's old number in order to aid

accurate retrieval of the new number. To evaluate the inhibitory account of RIF, Groome and Grant (2005) had participants complete both a RIF task and the *Cognitive Failures Questionnaire* (CFQ; Broadbent et al., 1982). Groome and Grant's argument was that if cognitive inhibition is the primary mechanism driving RIF, and if it is in fact an adaptive mechanism, then individuals who demonstrate a greater degree of forgetting (i.e., more inhibition) should also demonstrate fewer cognitive failures. Groome and Grant's results supported the inhibition account of RIF by finding a significant, inverse correlation between RIF scores and total CFQ scores. Therefore, to examine convergent validity of the degree of forgetting produced through the RIF procedure, participants in the current study were asked to complete the 25 item CFQ (Broadbent et al., 1982; see Appendix E).

The CFQ is a self-report measure of the degree to which individuals experience absentmindedness (e.g., daydreaming), memory deficits (e.g., forgetting what one went to the store to purchase), or slips of action (e.g., accidentally throwing away an item one intended to keep). Beyond total scores reflecting a single dimension of general cognitive failure (Broadbent et al., 1982; Larson, Alderton, Neideffer & Underhill, 1997), the factor structure of the CFQ is debatable. Some researchers argue for a five factor structure (Pollina, Greene, Tunick & Puckett, 1992), while others argue for a four factor structure (Matthews, Coyle & Craig, 1990) with agreement on only two dimensions. More recent research however (Larson et al., 1997) examined the factor structure of the CFQ with 2,949 participants and compared the results to Pollina et al.'s (1992) and Matthews et al.'s (1990) reported structures. Cronbach's alpha in this sample was .92, similar to Broadbent et al.'s (1982) original estimate, with evidence for only a single factor solution.

3.3.4. Discriminant validity materials. A 17 item self-report measure of social

desirability called the *Social Desirability Scale – 17* (SDS-17; Stöber, 2001; see Appendix F) was used to evaluate discriminant validity. The SDS-17 has demonstrated acceptable reliability and validity. For example, convergent validity correlations range from .52 - .85 with other social desirability measures, and substantial correlations with the Marlowe-Crowne Scale have also been obtained (Stöber, 2001). The SDS – 17 is suitable for use with adults between the ages of 18 – 80 years (Stöber, 2001).

3.3.5. Demographics. A brief, three question demographics questionnaire was created to collect participants' age, gender and first language (Appendix G).

3.3.6. Visual search task. During the 5 minutes of filler activity between retrieval-practice and final test of each RIF task, participants were asked to complete either a questionnaire (CFQ, SDS – 17, Demographics) trailed by a visual search task (Appendix H), or the visual search task alone. Assignment of which questionnaires were completed during each RIF task was random.

3.3.7. Study packages. All of the above-mentioned materials were organized according to the counterbalancing required to complete each task. These ordered materials were then placed in an envelope with a label affixed that indicated the order of materials inside and instructions for participants to create their "personal code" (Appendix I).

3.4. Procedure

Study participants were recruited through poster advertisements placed around the University of Saskatchewan campus, as well as various high public traffic areas and organizations within the community. The undergraduate Psychology Participant Pool was also used to recruit participants (https://usask.sona-systems.com/). All individuals who signed up to participate were tested. To aid participant retention across phases, the researcher sent

participants an email reminding them of their second testing date 1 – 3 days before the arranged time.

3.4.1. Phase I. Data collection occurred either individually or in small groups (2 – 3 participants). Participants who were tested individually were randomly assigned a prepared study package. Participants tested in groups were randomly assigned one of the pre-selected packages that contained the same RIF task orders (e.g., word list RIF task first, followed by facts) but different study, retrieval-practice and/or test orders to allow group collection to occur.

Following informed consent (Appendix J) participants were provided with either a word list or facts RIF task instruction booklet that contained instructions for each step of the procedure on its own page (refer to Appendix K and L for the word list and facts instructions, respectively). The first page of instructions was read aloud by the researcher while participants followed along on their own copies. Once all questions were addressed, the first RIF task began. Each RIF task commenced with participants flipping over their assigned study book and beginning to study the item presented on the first page. Once 5 seconds had passed, a computerized metronome (www.webmetronome.com) made an audible 'tick' indicating to participants that they should turn to the next page and begin studying the next pairing. After all items had been studied once, the researcher retrieved the study booklet(s) and provided participants with their assigned retrieval-practice booklet, face down in front of them. Participants were then asked to turn to the second page of instructions that were read aloud by the researcher and described the retrieval-practice task that they were asked to complete. Any questions were addressed and then participants were instructed to "go ahead" and begin completing their retrieval-practice booklet. Participants placed their retrieval-practice booklets into their study package envelope once they had finished and the researcher began a 5 minute timer. Participants were then asked to fill in

their personal code on the front of their study package and once complete, begin answering their first randomly assigned questionnaire or visual search task. If participants were assigned a questionnaire and completed it before 5 minutes had passed, they were instructed to place the questionnaire in their envelope and to begin completing the visual search task. Those who were assigned the visual search task worked on the two page form for the entire 5 minutes.

Once the filler activity was over, participants were instructed to turn to the final page of the instruction booklet and once again, the researcher read the instructions out loud. Participants were then provided with their recall booklet and were instructed to "go ahead" and begin recalling as many words or facts as they could that they remembered studying with each category cue during the initial study task. Participants placed their recall booklet inside their envelope when they were finished their recall.

This general procedure of study, retrieval-practice, filler activity, and final test was repeated with participants using their second assigned RIF task materials. Phase I testing took 30 minutes to complete. Participants from the Psychology 110 participant pool at the University of Saskatchewan were provided with one credit for their Phase I participation while community participants were provided with a $5.00 gift of thanks.

3.4.2. Phase II. Thirteen to fifteen days later, participants returned to the laboratory to complete Phase II. In the instances when a participant contacted the researcher to reschedule the Phase II time, accommodations were made to test the participant at the earliest possible time. Testing occurred either individually or in pairs.

Informed consent was briefly reviewed and participants were asked to provide the first three letters of their Mother's first name and the day and month of their birth to locate their particular study package. Two more RIF tasks were then completed using participants' assigned

materials. The steps and instructions used for the RIF tasks completed at Phase I were identical for those completed at Phase II, with the exception of the materials used (i.e., facts or words). Participants were then debriefed (Appendix M) and thanked for their time. Undergraduate Psychology Participant Pool participants were granted another course credit for their participation while community participants received another $5.00 gift of thanks.

3.4.3. Phase III. Nineteen to 55 days later, all participants were invited via email to return to the lab one final time to compete the RIF task using facts again. Eighteen participants chose to complete the final phase of the study. The task followed exactly the same procedure and counterbalancing orders used for the first facts RIF collection with the exception of the visual search task being the only form administered during the filler activity. Once finished, participants placed their materials into their envelopes and were provided with a final $5.00 gift of thanks.

3.5. Coding

Recall data from each of the five RIF tasks that participants completed were coded according to the three practice types used in RIF tasks, Rp+, Rp- and NRp. The number of words or facts that participants accurately recalled per practice type was totalled and converted into a proportion. Misspellings of words were disregarded and scored as accurate if the intended word was obvious (e.g., dafodill = "daffodil"). For the facts RIF tasks, participants need not to have written down the sentence in its entirety for it to be scored as accurate recall, but the ideas conveyed in the original sentence had to be obvious. For example, the fact "There are 260 different varieties of spider on Bilu," would be scored as correct if the participant wrote down "They have 260 different kinds of spiders," under the heading of "Bilu" but incorrect if the participant wrote down "260" or "spiders."

Prior to examining the psychometric properties of RIF, typical RIF effects had to be demonstrated in all tasks. Had typical RIF effects not been detected, it would be reasonable to conclude that there was a confound within the materials or procedure and therefore examining the reliability of the effect would then be unwarranted. Once typical RIF effects were demonstrated for each task (i.e., significantly higher recall of Rp+ details and significantly lower recall of Rp- details when compared to the NRp baseline), *degree of forgetting* (DOF) scores were calculated for each RIF task that participants completed. To calculate the DOF scores, the proportion of NRp details were subtracted from the proportion of Rp- details recalled (Anderson et al., 1994). DOF scores are typically negative, with more negative numbers indicating a greater degree of forgetting.

4. CHAPTER 4: RESULTS

Screening for missing values and evaluating the statistical assumptions that must be met in order to evaluate the study hypotheses are provided first in this chapter, followed by a detailed description of the study results.

4.1. Data Cleaning and Preliminary Analyses

Prior to evaluating the hypotheses of the current research, data were screened for missing values and issues concerning the normality of the data (e.g., outliers, skewness, Kurtosis) that may affect the analysis and/or interpretation of the results. Next, to ensure that typical RIF effects had been obtained prior to evaluating the hypotheses of the current research, the proportion of details recalled from each practice type was entered into a repeated measure analysis of variance (RM ANOVA) for each RIF task completed.

4.1.1. Missing values. No missing values were obtained other than from those participants who withdrew from the study after the end of Phase I or Phase II testing. To help prevent missing data, the researcher reviewed participants' questionnaire responses between their Phase I and Phase II testing dates. If a response was missing, the researcher asked participants at their next meeting whether they were comfortable filling in a response, and offered them a chance to complete it. All participants indicated that they simply missed the question and filled in a response.

4.1.2. Statistical assumptions. Each of the three levels of practice (Rp+, Rp-, NRp) for each of the five RIF tasks as well as total scores on the CFQ and SDS – 17 were checked for outliers and normal distribution of the data using the Kolmogorov-Smirnov tests of normality. No outliers were found with all z-scores < 1.96 (Field, 2009). The distribution of scores for the proportion of Rp- words recalled during the first administration of the repeated word list RIF

task, the proportion of NRp words recalled during the second administration of repeated word list RIF task, and the proportion of NRp words recalled for the set of equated word list RIF tasks were the only variables that did not violate the assumption of normality, all *D*'s (56) < .12, *p*'s > .06. Kolmogorov-Smirnov test of normality for the remaining practice type variables, total CFQ scores and SDS – 17 scores were all statistically significant, all *D*'s > .12, all *p*'s < .03.

4.2. Retrieval-Induced Forgetting

One RM ANOVA using the three levels of practice (Rp+, Rp-, NRp) was performed on each RIF task separately to determine whether or not typical RIF had been obtained. Main effects were then followed up by paired-samples t-tests. Given that normality had been violated for a number of the levels of practice variables, all RM ANOVAs were run again using Friedman's related samples ANOVA for nonparametric data (Field, 2009). Post-hoc *t*-tests were also run again using Wilcoxon's signed-rank test with the Bonferroni correction applied to control for Type I error when conducting nonparametric post-hocs (Field, 2009). The assumption of sphericity was also violated for all variables included in the RM ANOVAs. However, using these nonparametric tests and examining two adjustments to address the violation of sphericity (Greenhouse-Geisser correction, Huynh-Feldt correction) did not change the pattern or statistical significance of the results so only the parametric data analyses are reported. Results for each RIF task is discussed in turn next.

4.2.1. Retrieval-practice success. The average retrieval-practice success rate for all RIF tasks was consistent with past research for both category-word pairs (e.g., Anderson et al., 1994) and sentences (e.g., Macrae & McLeod, 2001). For the category-word pair RIF tasks that participants completed twice, the average retrieval-practice success rate was 89% (Phase I) and 96% (Phase II). Retrieval-practice success was also high for the equated category-word pair RIF

task with an average success rate of 87%. Finally, average retrieval-practice success rates of 86% and 89% were obtained for the first and second administration of the facts RIF task, respectively.

4.2.2. Test-retest repeated words. For the category-word pair RIF task that participants completed twice, 2 weeks apart, typical RIF effects were obtained. For Phase I, a one-way RM ANOVA revealed a main effect of practice, F (2, 98) = 109.27, $p < .001$, $\eta_p^2 = .69$. Paired-samples t-tests indicated that participants recalled significantly more Rp+ words ($M = .76$, $SD = .17$, 95% CI [.73, .82]) than Rp- ($M = .40$, $SD = .22$, 95% CI [.35, .46]) and NRp ($M = .49$, $SD = .15$, 95% CI [.45, .53]) words, t (49) = 12.52, $p < .001$, and t (49) = 11.63, $p < .001$, respectively. Significantly fewer Rp- words were recalled than NRp words, t (49) = -3.73, $p = .001$.

For the RIF task using the same words at Phase II, a main effect of practice was also obtained, F (2, 98) = 88.23, $p < .001$, $\eta_p^2 = .64$, with paired-samples t-tests indicating significantly more Rp+ words ($M = .86$, $SD = .13$, 95% CI [.81, .88]) were recalled than Rp- ($M = .51$, $SD = .24$, 95% CI [.46, .58]), t (49) = 10.83, $p < .001$, and NRp ($M = .60$, $SD = .17$, 95% CI [.54, .64]) words, t (49) = 10.53, $p < .001$. Significantly fewer Rp- words were recalled than NRp words, t (49) = -3.79, $p < .001$.

4.2.3. Test-retest equated words. For the set of words that were equated to the set that participants completed twice, typical RIF effects were also obtained. A one-way RM ANOVA revealed a main effect of practice, F (2, 98) = 97.58, $p < .001$, $\eta_p^2 = .67$. Paired-samples t-tests indicated that participants recalled significantly more Rp+ words ($M = .74$, $SD = .16$, 95% CI [.71, .79]) than Rp- ($M = .41$, $SD = .23$, 95% CI [.35, .47]) and NRp ($M = .49$, $SD = .17$, 95% CI [.45, .54]) words, t (49) = 11.40, $p < .001$, and t (49) = 11.15, $p < .001$, respectively. Significantly fewer Rp- words were recalled than NRp words, t (49) = -3.61, $p = .001$.

4.2.4. Facts. For the set of facts presented in sentence format the first time, typical RIF effects were obtained. A one-way RM ANOVA demonstrated a main effect of practice, F (2, 98) = 92.71, $p < .001$, $\eta_p^2 = .65$. Paired-samples t-tests indicated that participants recalled significantly more Rp+ facts ($M = .62$, $SD = .23$, 95% CI [.56, .68]) than Rp- ($M = .14$, $SD = .18$, 95% CI [.10, .20]) and NRp ($M = .51$, $SD = .30$, 95% CI [.43, .61]) facts, t (49) = 13.06, $p < .001$, and t (49) = 2.64, $p = .011$, respectively. Significantly fewer Rp- facts were recalled than NRp facts, t (49) = -11.18, $p < .001$.

Another RM ANOVA was run for the 16 participants who completed the facts RIF task a second time, and standard RIF effects were also obtained, F (2, 34) = 67.05, $p < .001$, $\eta_p^2 = .80$. Paired-samples t-tests revealed that significantly more Rp+ facts ($M = .64$, $SD = .19$, 95% CI [.55, .74]) were recalled than both Rp- facts ($M = .14$, $SD = .20$, 95% CI [.04, .25]) and NRp facts ($M = .22$, $SD = .14$, 95% CI [.15, .29]), t (17) = 8.50, $p < .001$, and t (17) = 9.50, $p < .001$ respectively. The proportion of Rp- facts recalled was also significantly lower than the proportion of NRp details recalled, t (17) = -2.23, $p = .039$.

Each RM ANOVA mentioned above was performed again with individuals who did not speak English as their first language excluded from the analysis. The pattern of results remained the same, thus these participants were not excluded from subsequent analyses. Similar ANOVAs were also run with gender as a between-subjects variable which produced no significant gender × practice type interactions (all p's > .62). Refer to Figure 4.1 for a visual depiction of RIF across all tasks.

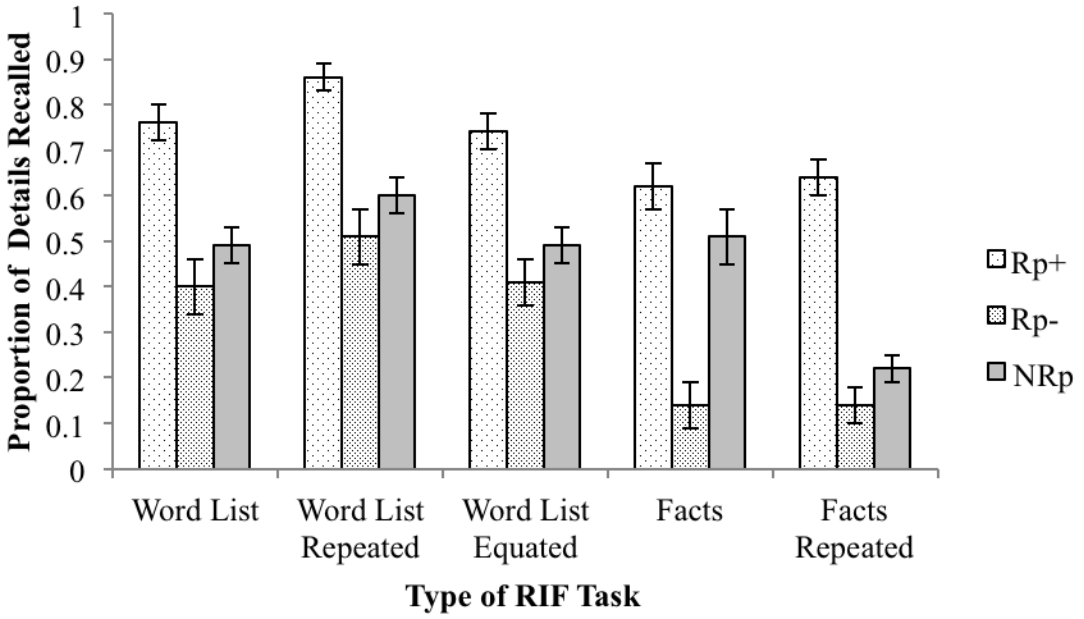


Figure 4.1
The proportion of details recalled for each practice type across all RIF tasks. Error bars represent the standard error of the means.

4.3. Psychometric Properties of Retrieval-Induced Forgetting

Degree of forgetting (DOF) scores for each RIF task were calculated by subtracting the proportion of NRp details recalled from the proportion of Rp- details recalled for each RIF task. DOF scores are typically negative with more negative numbers indicating a greater degree of forgetting (Anderson et al., 1994). Refer to Table 4.1 for mean DOF scores and standard deviations across all RIF tasks.

RIF Task	Range	Mean (SD)
DOF Test-Retest Words Phase I	-.39 to .42	-.09 (.17)
DOF Test-Retest Words Phase II	-.42 to .25	-.09 (.17)
DOF Test-Retest Other Words	-.36 to .25	-.08 (.15)
DOF Facts	-.80 to .40	-.38 (.24)
DOF Test-Retest Facts	-.20 to .40	-.08 (.15)

Table 4.1
Descriptive statistics for degree of forgetting scores across the five retrieval-induced forgetting tasks.

4.3.1. Delay between Phase I and Phase II testing. The average delay between Phase I and Phase II testing times was 14 days (*SD* = 1.72; *range*: 6 – 18 days). To determine whether or not the length of delay between testing phases impacted DOF scores for the category-word RIF task that participants completed twice, a bivariate correlation was performed. A significant negative correlation was found between the length of delay and DOF scores for repeated category-word pairs, r (50) = -.31, p = .027.

4.3.2. Delay between Phase II and Phase III testing. Participants were invited via email to complete a brief follow-up testing session (Phase III). The average length of time between Phase II and Phase III testing was 36 days (*SD* = 10.48, *range:* 19 – 55 days). A bivariate correlation was performed between Phase III DOF scores and the length of delay between testing. No significant correlation was found, r (18) = -.08, p = .766.

DOF scores for each RIF task, CFQ scores and SDS – 17 scores were then entered into a bivariate correlation matrix to examine test-retest reliability, alternate forms reliability, convergent validity and discriminant validity evidence. The results regarding each psychometric property are discussed in turn next.

4.3.3. Word list test-retest reliability. A significant positive correlation was found between DOF scores for the test-retest words that participants completed at both Phase I and Phase II, r (50) = .46, p = .001. No significant correlations were found between DOF scores for the test-retest repeated words and the equated category-word RIF tasks at both Phase I, r (50) = .05, p = .72 or at Phase II, r (50) = .02, p = .89. A partial correlation was performed controlling for the length of delay between Phase I and Phase II testing to evaluate whether or not the test-retest reliability coefficient would persist. Even after controlling for the length of delay between repeating the category-word RIF task, a significant positive correlation was obtained, r (50) =

.42, p = .003, accounting for 17.6% of the variance.

4.3.4. Facts test-retest reliability. A significant positive correlation was found between DOF scores for the two administrations of the facts RIF task, r (18) = .51, p = .032, accounting for 27% of the variance.

4.3.5. Alternate forms reliability. DOF scores between the facts RIF task and the three category-word RIF tasks were correlated to examine alternate-forms reliability evidence. All correlations were non-significant. The correlation between the facts DOF scores and the test-retest words completed at Phase I was non-significant, r (50) = -.26, p = .071, as was the correlation between facts DOF scores and the test-retest words completed at Phase II, r (50) = -.10, p = .494, and the set of equated test-retest words, r (50) = -.06, p = .686. DOF scores for the second administration of the facts RIF task also did not significantly correlate with any word list DOF scores, all p's > .42.

4.3.6. Convergent validity. Total CFQ scores (M = 45.11, SD = 9.68, 95% CI [42.52, 47.70] were entered into a bivariate correlation matrix along with DOF scores from all RIF tasks. No significant correlations emerged for either the test-retest words completed at Phase I or at Phase II, r (50) = .05, p = .723, and r (50) = .12, p = .427, respectively. The correlation between total CFQ scores and DOF scores for the set of equated test-retest words, r (50) = .05, p = .713, and for facts DOF scores during the first and second administrations were also non-significant, r (50) = -.05, p = .747, and r (18) = .38, p = .122, respectively. Given the mixed results regarding the factor structure of the CFQ (Broadbent et al., 1982; Larson et al., 1997; Matthews et al., 1990; Pollina et al., 1992), scores for the various reported subscales were also calculated and correlated with DOF scores. Again, no significant correlations emerged, and all p values were greater than .117.

4.3.7. Discriminant validity. Evidence for discriminant validity was obtained through non-significant correlations between DOF scores and scores on the SDS-17 (M = 6.60, SD = 2.74, 95% CI [5.92, 7.28]). No significant correlations were obtained for the category-word RIF task that participants completed at Phase I, r (50) = .04, p = .777, as well as at Phase II, r (50) = .12, p = .427. SDS-17 scores also did not correlate with the equated word list RIF task, r (50) = .17, p = .25, or the facts, r (50) = .19, p = .182.

5. CHAPTER 5: DISCUSSION

This final chapter first provides the reader with a review of the purpose and importance of the current study and then moves into an interpretation of the study results. Potential limitations and suggestions for future directions are also provided along with the conclusions that can be drawn based on the current results.

5.1. Review and Interpretation of the Study Results

Despite the importance of establishing the psychometric properties of assessment tools, many research-based measures are frequently used, and the scores interpreted, without such evidence available. Without an understanding and evaluation of the psychometric properties of scores obtained from a measure, one cannot be sure of what the measure is assessing, or whether or not the assessment is stable, or reliable. Use and interpretation of scores that lack reliability has great potential to lead to invalid use of the scores through misapplication and/or misinterpretation of results. A goal of psychological and educational assessment is to obtain scores that reflect a measurement of the construct of interest and to use those scores in some form of decision making (Crocker & Algina, 1986; DeVellis, 2003). For example, scores from an intelligence test may be used to make admission, placement, or diagnostic decisions (Crocker & Algina, 1986). If such decisions were made based on unreliable intelligence test scores, negative outcomes would likely ensue as changes in obtained scores would not be due to a measured change in the construct of interest but rather due to changes in error. In such instances, individuals with intelligence deficits may go without intervention, poorly qualified individuals may be admitted into a program that is too advanced for them, while other individuals may be erroneously diagnosed with an intelligence deficit. Regardless of the decision to be made from test scores, valid application of those scores can only be made when the obtained scores are

reliable (Cohen et al., 2007).

There are instances when there is a lack of psychometric evidence available to help users of the measure confidently interpret the obtained scores, yet the measure remains in use. Although continued use of a measure that lacks psychometric evidence may further theory, conclusions and applications of the scores must be cautiously made as a clear understanding of exactly what the measured construct is or how stable it may be has not been empirically evaluated. Without reliability evidence, the decisions made based on scores obtained during one administration of the measure may be very different from decisions made based on a later administration of the same measure. Or, scores on two forms of a measure may be different from one another not because of differences in the measured construct but because of the poor reliability of the measures (not measuring the same construct as intended). An example of this potentially invalid interpretation and application of scores can be found in the literature surrounding RIF. The tendency to temporarily forget unretrieved information from memory following repeated retrieval of related information, or RIF effects, appears to be robust throughout the literature in that the same pattern of results has been obtained using a variety of materials (e.g., word lists, Anderson et al., 1994; facts, Macrae & MacLeod, 1999; social cognition, Macrae & MacLeod, 1999; autobiographical memory, Barnier et al., 2001; eyewitness memory, Shaw et al., 1995; visuospatial memory, Ciranni & Shimamura, 1999) across different populations (e.g., children, Ford et al., 2004; adults, Migueles & García-Bajos, 2007; clinical populations, Nestor et al., 2005). Regardless of such intense research interest, very little is known regarding the validity and reliability of the forgetting scores produced through the procedure leading to reduced confidence in decisions and interpretations made based on those scores. For example, once RIF has been found using a certain set of materials, further analyses

may be performed by using groups of individuals who have been classified as being either high or low on the construct of interest (e.g., those who demonstrate a high degree of forgetting versus those with who demonstrate no or very little forgetting). Without knowing how reliable the scores are in measuring the underlying construct, conclusions may be drawn based on data that is composed of more error, or unaccounted variance, than true score variance. Decisions and interpretations made from one administration of the measure may be very different during a different administration of the measure. The purpose of the current research was to empirically evaluate the psychometric properties of the forgetting scores produced through the RIF procedure in order to inform researchers and practitioners about the stability of RIF scores across time and materials. Validity evidence was also sought through the current research with the goal of providing enough psychometric information to increase confidence in RIF score use and interpretation. To accomplish this, participants were asked to complete three RIF tasks using category – word pairs and two RIF tasks using facts in sentence format. Forgetting scores from each RIF task were then correlated with each other and with scores obtained on a social desirability, and cognitive failures questionnaire.

Retrieval-induced forgetting effects (i.e., significantly lower recall for Rp- details than both Rp+ and NRp) were obtained for all RIF tasks however, counter to theoretical prediction, evidence of test-retest reliability was obtained only when identical materials were used. Strong positive correlations emerged between the category – word pair RIF tasks that participants completed twice as well as between the two administrations of the facts RIF task. The length of time between the two administrations of the word list RIF task influenced the strength of the relationship found between the repeated word list RIF tasks, indicating that the 2 week testing interval was not long enough to allow all carry-over effects to dissipate. When the impact of

delay was controlled in the word list RIF correlation, a significant positive relationship was still observed suggesting that participants' true scores for that construct were being measured and that the results were not simply due to carry-over effects. Further evidence that true scores were measured and that the results are not purely due to carry-over effects was found in the facts RIF task correlation. The facts RIF tasks were separated by a longer delay than the repeated word list RIF tasks which should have given more time to eliminate carry-over effects. Regardless of the longer delay, a strong positive correlation was still obtained between the two administrations of the facts RIF tasks. All other correlations with RIF scores were non-significant.

The overall results of this research can be more easily understood using the target example from Chapter 2 that was used to illustrate the differences in reliability and validity. Recall that the goal of assessment is to hit the center of the target (the construct of interest) a number of times (stability). If such reliability is obtained, then users of the measure will have more confidence in the scores and will be more likely to make valid applications of those scores. Adding to that example, also imagine that there are two different kinds of darts that can be thrown at the target. The two different darts are considered to be identical to one another with only one difference – their color – making the two types of darts alternate forms of each other. If the different coloured darts measure the same construct in a consistent manner, then both colours should repeatedly hit the bulls-eye center of the target (alternate forms and stability reliability). If, however, there are unknown differences in the two different coloured darts, when thrown at the target, the darts may repeatedly hit the same location (stability reliability) but two clusters of the same coloured darts will be found at different locations on the target (i.e., the darts are grouped by colour around two separate spots on the target). In this instance, the two different measures (darts) would be argued to have good stability reliability but the measures would not

demonstrate alternate forms reliability. The clusters of darts would appear to be tapping different constructs, or the same/similar construct that is sensitive to factors that remain unaccounted. Using this example with the current results, stability reliability of RIF scores was found by repeatedly hitting the same spot on the target when using the same materials (same colour of darts). Alternate forms reliability was not demonstrated however, as scores from the two different forms of materials, words and facts (two different coloured darts), did not correlate with one another (did not hit the same location on the target). Taking this example one step further, now imagine that a new box of darts has been ordered that are advertised as being cheaper to purchase but otherwise identical to one of the colours of the previous darts used. When these new darts are thrown at the target however, they consistently hit the same spot on the target but it is not the same area that the old darts of that colour typically hit. These findings would lead one to conclude that the new darts are demonstrating stability reliability but also that there must be differences in the two kinds of darts, or in the sensitivity of the darts to some unknown factor(s), that the manufacturer did not anticipate. In the current research, no correlation was obtained between the two sets of category – word pairs that were matched according to taxonomic frequency (the identical old and new darts). It is possible that the two sets were not matched to each other on certain unknown factors that contribute to a significant amount of the variance in RIF scores (some unknown differences between the old and new darts). It may also be that the underlying construct, or driving force of RIF scores, is more sensitive to situational factors and/or task demands than initially thought. Discussion of these two potential explanations of the current results is provided next.

The lack of alternate forms reliability evidence when correlating the word list and facts RIF tasks, or more surprisingly, test-retest reliability for the equated materials suggests that the

degree of forgetting that participants demonstrate through a RIF procedure is likely influenced by more than simply inhibitory mechanisms and taxonomic frequency of the items within categories (Anderson et al., 1994). Indeed, a few boundary conditions to RIF have been identified that reduce or eliminate RIF effects. Inserting a long delay between retrieval-practice and final test (MacLeod & Macrae, 2001) or using different final recall procedures (Butler, Williams, Zacks & Maki, 2001) have been reported to moderate the forgetting found in the typical RIF procedure while detailed integration of items during encoding has also been shown to reduce the degree of forgetting obtained (Anderson & McColloch, 1999). Some researchers have argued that self-relevant information is resistant to RIF (Macrae & Roseveare, 2002) while others have found temporary forgetting of autobiographical information with the procedure (e.g., Barnier et al., 2001; Harris et al., 2010; Marche et al., 2011; Wessel & Hauer, 2006), with individuals' mood (Harris et al., 2010; Rhyno, 2008) and the emotional valence of the materials (Barnier et al., 2001; Harris et al., 2010) moderating the degree of RIF obtained. The current results, combined with our understanding of the various boundary conditions of RIF, suggest that the materials and both inter- and intra-individual differences impact RIF scores. Therefore, researchers will need to carefully interpret their RIF findings with an understanding that the materials used, and possibly other currently unknown factors, are likely impacting the degree of RIF obtained. Individual differences in unintentional forgetting may be measured through the RIF task, but interpretations will also need to include considerations of the materials used, potential boundary conditions, and the degree to which they may have impacted the results.

Strong positive correlations between scores obtained from two administrations of the same measure is interpreted as test-retest reliability evidence – although there is variance in the scores, variability in one score is met with similar variability in the second score (Crocker &

Algina, 1986). Error from changes in participants' state, the measurement procedures used, or other random error may slightly impact the results but participants' true scores for the latent variable are being consistently tapped across the two testing sessions. Evidence of this form of test-retest reliability was attained in the current research by finding positive correlations between repeated materials. Carry-over effects were expected to influence this relationship due to participants' previous experience with the materials, thus an equated set of materials was designed to provide a test-retest reliability estimate with the variance due to carry-over effects removed. Positive correlations that were slightly lower than the correlation between the repeated materials were expected for scores obtained from the equated RIF task materials however quite surprisingly, no relationship was found. These findings suggest that the 'equated' materials were not necessarily equated according to certain factor(s) that account for a significant amount of variance in the scores. There may also be unknown individual or situational factors that are impacting the expected relationship between scores or a combination of these potential confounds may be at play – this cannot be discerned through the present research. What can be concluded however is that RIF scores appear to demonstrate temporal stability when the same materials are used and that the factors that contribute a significant amount of variance to RIF scores remain unclear. Research regarding boundary conditions of RIF (e.g., Anderson & McColluch, 1999; Bäuml & Kuhbander, 2007) should continue to help elucidate the factors that contribute to, or hinder, RIF.

Anecdotal evidence from participants following debriefing provided some additional support for the notion that other factors are contributing to the degree of RIF obtained across tasks. Certain items, or entire categories from a specific RIF task, may be more or less memorable or salient to some individuals leading to a differential impact of retrieval-practice for

these items/categories. For example, following debriefing, many participants chose to verbally share their personal memory strategies or efforts made to remember as many items as possible across each RIF task. One participant indicated that once she was aware that more than one item was going to be paired with each category, she mentally counted the number of items from a category during study, and worked at recall until she reached that number of items per category. Although typical RIF effects may be obtained for this participant across all RIF tasks, the degree of RIF displayed will likely differ depending on when (during which RIF task) the participant decided to count items within a category. This extra effort at recall may overcome the inhibition produced through the mental competition of Rp+ and Rp- items leading to a slightly elevated level of Rp- recall. The idiosyncratic memorability of certain items or entire categories may also contribute to the degree of RIF obtained. If one or more categories of items are more memorable or salient to the self than others, it is likely that those items or categories will be more resistant to forgetting (Macrae & Rosenveare, 2002). After debriefing, another participant indicated that as a "car buff" she works on different vehicles almost daily and therefore knows a great number of car manufacturers. She claimed that no effort was required in order to encode and recall all of the items that were paired with the "Car" category due to her extensive experience with cars. With these two examples, it is easy to see how the degree of RIF obtained from participants may be impacted by differences in materials that are beyond taxonomic frequency, in addition to being impacted by participants' level of cognitive inhibition. Examining the amount of shared variance between a measure of cognitive inhibition, such as the Stroop task (e.g., MacLeod, 1991), and a variety of RIF tasks would help determine the extent to which mental inhibitory mechanisms impact RIF scores. However, similar to research on RIF and the DRM paradigm, the psychometric properties of the Stroop task have not been fully developed (e.g., Kindt,

Bierman & Brosschot, 1996) leading to an incomplete understanding of exactly what inhibitory mechanisms are being measured by the Stroop task. Further research examining correlates of RIF scores is needed in order to continue defining the mechanisms impacting RIF scores and subsequently developing RIF theory.

The absence of alternate forms reliability and support for test-retest reliability makes interpretations regarding the validity of RIF scores as a measure of individual differences in forgetting ability rather difficult. Reliability is a necessary precondition to validity (Cohen, Manion & Morrison, 2007) – if obtained scores are not replicable, or reliably obtained, then valid and confident use of those scores in decision-making is next to impossible. Obtaining reliable scores however, does not guarantee valid interpretation and use of those scores (Cohen et al., 2007). Evaluating the discriminant and convergent validity evidence of RIF forgetting scores as a measure of individual differences in forgetting ability would be a rather insignificant contribution at this juncture. No correlation between RIF scores and SDS – 17 scores was predicted which would provide evidence to indicate that two different constructs were being measured and the current results support this claim. However, such a conclusion is not especially informative given that a significant amount of variance in RIF scores cannot yet be predicted. We could accurately predict that the pattern of results following a RIF procedure would be similar with a variety of materials (i.e., typical RIF effects) but the degree, or amount of forgetting, that would occur when using those materials could not yet be predicted. Researchers in the future may attempt to create RIF score norms for a particular set of materials that is devoid of known boundary conditions (e.g., pronounceable nonwords) in order to compare different groups' ability to forget through the procedure with the same materials. However, exactly what construct(s) is/are being measured with those materials and how to appropriately

interpret and apply the results could not yet be discerned as we still cannot reliably account for a significant amount of the variance in forgetting scores across RIF tasks. The current results suggest that if such materials were developed, it is likely that test-retest reliability would be obtained between multiple administrations of the same materials. Classical test theory would then lead one to argue that this test-retest reliability evidence supports the notion that individuals' *true scores* for the same construct were being measured. However, valid interpretations and use of the scores at that point could still not occur as the construct(s) that the true scores represented would still be unknown. Creating such a set of materials may be a start towards finding a more pure measure of participants' true scores for the construct(s) underlying RIF scores. However, without continuing that line of research by finding factors that account for a significant amount of variability in different RIF scores, we still could not conclude with confidence that the latent variable being measured by that particular set of materials is the same latent variable that is being measured with other RIF materials.

It also remains quite possible that the same underlying construct is driving forgetting scores across different RIF tasks, but that the mechanism is differentially impacted by the context of measurement. For example, it may be that certain types of cognitive inhibition account for a large amount of variance in only certain RIF scores that are obtained using a specific set of materials. Or perhaps there are other factors that are influencing the levels of measured cognitive inhibition across materials and/or within individuals that in turn impact RIF scores (e.g., working memory capacity, Aslan & Bäuml, 2011). Great confidence cannot be placed in either of these two explanations, nor can great confidence be placed in any other explanation or combination of explanations – the RIF literature has not provided enough psychometric evidence to warrant great confidence that valid score use occurred.

Assessing convergent validity in the current research is also problematic. Past research had found an inverse correlation between RIF scores and scores obtained from the CFQ (Broadbent et al., 1982) that the researchers argued provided support for the inhibitory account of the effect (Groome & Grant, 2005). No such correlation emerged in the current research even when subscale scores of the CFQ were used in the analysis. The current research obtained only test-retest reliability evidence for identical materials (rather than the equated materials and alternate form materials) therefore drawing conclusions about convergent validity of the CFQ and RIF scores remains tenuous. Groome and Grant's (2005) results may be replicable, but perhaps only when the same or similar materials to those that they used are employed for the RIF task. Groome and Grant's research made use of category – word pairs, as did the current research, however only 36 pairs were initially studied while in the present study 60 pairs were studied. It is possible that the CFQ shares an inverse relationship with RIF scores but only for shorter RIF tasks that will not overload working memory (Aslan & Bäuml, 2011). Or perhaps RIF scores are quite sensitive to task instructions or other unknown factors. Again, conclusions cannot be drawn with confidence as we still cannot accurately predict an adequate amount of variance in RIF scores across tasks and within the same individuals.

A second possible explanation for the failure to replicate Groome and Grant's (2005) research is that the CFQ itself is not a reliable measure of 'forgetfulness.' Research using the CFQ has unfurled since the early 1980's yet consensus, or even reasonable agreement, regarding the factor structure of the CFQ has not yet been reached (e.g., Broadbent et al., 1982; Larson et al., 1997; Matthews et al., 1990; Pollina et al., 1992). The interpretations and inferences that researchers make based on their results are only as strong as the measurement tool used to assess the construct(s). If the CFQ is not reliably tapping 'forgetfulness' (unintentional or otherwise),

then invalid interpretations and applications of the results may occur. A Type I error may have occurred in Groome and Grant's study; a Type II error may have occurred in the current study or once again, there may be other unknown factors influencing the results. Future research may be directed toward disentangling the relationship between CFQ and RIF scores using a variety of RIF materials in attempt to make a definitive conclusion about if, and when, the relationship exists. Including other measures of cognitive failure and/or inhibition in such research would also help to inform theory and define the type of forgetting that occurs through the RIF procedure. By correlating RIF scores obtained from using different materials to scores obtained from other known measures of cognitive function and/or failure, the results would provide a means of further assessing the validity of the CFQ as a predictor of RIF scores. Strong conclusions regarding the potential relationship between CFQ and RIF scores cannot be drawn until further research attempts to replicate the current results or those of Groome and Grant.

5.2. Limitations and Future Directions

The finding of test-retest reliability of RIF scores using the same materials has a number of implications for theory and research regarding RIF. After data collection for the current study was complete, one in press study was found that examined the test-retest reliability of word list RIF scores using a variety of final test procedures (e.g., cued recall, category – stem recall, recognition; Potts, Law, Golding & Groome, in press). In line with the current research, Potts et al.'s (in press) results closely mirror those found in the current study – test-retest reliability was only obtained when identical materials were used. Obtaining similar results from independent researchers lends empirical support to these novel findings. The current research and Potts et al.'s findings have identified the need to carefully interpret RIF results especially when making comparisons across materials and/or individuals. Individuals may get different scores from one

another but the difference may be due to other factors influencing participants' true scores rather than a measured difference in the trait of interest. An examination of the 95% confidence intervals further suggests that participants' true scores are being differentially impacted within individuals (when different materials are used) rather than across individuals (as would be expected with a reliable individual difference measure). Estimates of where participants' true scores lay are not very similar within participants or across materials and the range of the interval could be considered quite wide. These observations suggest that the degree of RIF obtained for one task is not necessarily the same (or similar) degree of RIF that would be found for another RIF task completed by the same individual.

The current study is not without its limitations. Some may argue that participants' experience with earlier RIF tasks and/or their knowledge of the effect may have impacted the results obtained. During the first two testing sessions, participants completed two RIF tasks at each meeting and were then debriefed. Some participants chose to come back to the laboratory to complete Phase III testing following debriefing. Upon initial consideration, one may argue that only the first RIF task that participants completed would be a measure of RIF without carry-over effects – the typical pattern of *study* → *retrieval-practice* → *distractor task* → *recall* was followed for each RIF task which may lead participants to try different memory strategies as the trials progressed. If such experience impacted participants' scores however, then significant differences and correlations should have only emerged for the first RIF task that participants completed. This was not the case however, as typical RIF effects were found for all tasks and, regardless of the order of completion of the tasks, significant positive correlations were obtained for repeated RIF tasks. Participants may have made extra or altered efforts to reduce forgetting because of their awareness of the task and/or effect but such efforts do not appear to have

confounded the results. In fact, a number of participants who returned to complete Phase III commented that they "tried really hard not to forget" items because they were aware of the effect from our debriefing discussion. Consistent with the current findings, past research has demonstrated that warning participants about RIF fails to reduce or eliminate the effect (Jones, 2010).

A final limitation of the research relates to the generalizability of the results. Autobiographical memory has been shown to both succumb to RIF (e.g., Barnier et al., 2004; Marche et al., 2011) and be resistant to RIF (Macrae & Roseveare, 2002) but an autobiographical RIF task was not included in the current research. Simple word list RIF tasks and a facts RIF task were selected as the materials to be used in order to develop two sets that could be matched according to taxonomic frequency and to examine a simple alternate form of the procedure using different materials. Test-retest reliability for the repeated materials was obtained in the current research, but a different pattern of results may have emerged had autobiographical materials been used. The inconsistent results in autobiographical RIF tasks may again be due to unknown boundary conditions, differential impact of the context of measurement on participants' true scores and/or error scores, or a combination of these factors. Other unknown factors may also impact autobiographical RIF scores – confident conclusions cannot be drawn yet as an examination of the psychometrics of autobiographical RIF scores has not occurred. If such a study were conducted, typical RIF effects would likely emerge, however in light of the current results, proffering further predictions regarding the stability or reliability of autobiographical RIF scores would be precipitous. Different materials did not share RIF score variance within individuals in the current study, and there are myriad factors that may contribute to autobiographical remembering. The current test-retest reliability results cannot be generalized to

autobiographical RIF tasks as the RIF scores obtained using different materials appear to be independent of one another so it would be reasonable to assume that autobiographical RIF scores would also differ. Future research may examine the psychometric properties of RIF scores obtained from autobiographical tasks to inform researchers who are moving towards application of the effect using more ecologically valid materials (e.g., eyewitness memory, García-Bajos, Migueles, & Anderson, 2008).

As research progresses towards a more comprehensive account of the factors that affect the variance found in RIF scores, valid applications and interpretations of the scores produced through the procedure will progress as well. With our current level of understanding, it can be concluded from the research that forgetting scores obtained from the RIF procedure are somewhat dependent upon the materials used, or the *context* of measurement (Crocker & Algina, 1986; Messick, 1989). Measures themselves are not considered valid or reliable, rather the scores obtained from the context of measurement are (Messick, 1989). Typically, when constructing a measure, questions or items are written to encompass all practical and theoretical domains of the construct as possible (DeVellis, 2003). The goal here is to ensure that all known facets of the construct are adequately represented in the measure. Similar to interview protocols, no explicit questions can be written to assess the incidental forgetting produced through the RIF procedure as researchers argue that the forgetting is due to an unconscious retrieval competition process (i.e., cognitive inhibition, Anderson et al., 1994). Regardless of the unique nature of the forgetting scores obtained through the RIF procedure, researchers continue to delineate RIF scores as a measure of incidental (or unintentional) forgetting. Before such research can be conducted with confidence, the factors that influence RIF scores (i.e., the different facets that make up the scores) must first be revealed. Cognitive inhibition may be one of the factors that

influence RIF scores, however the research literature is beginning to demonstrate that many more factors may also be in play. Valid conclusions, interpretations and applications of RIF forgetting scores can be made with confidence only once a significant amount of the variance in scores can be reliably accounted for. RIF effects remain robust across materials, but future research needs to determine the factors that influence individual differences in the degree of RIF obtained.

REFERENCES

Allen, M. J., & Yen, W. M. (1979). Introduction to measurement theory. Monterey, CA: Brooks/Cole Publishing Company.

Anderson, M. C. (2003). Rethinking interference theory: Executive control and the mechanisms of forgetting. *Journal of Memory and Language, 49* (4), 415 – 445. doi:10.1016/j.jml.2003.08.006

Anderson, M. C., Bjork, R. A., & Bjork, E. L. (1994). Remembering can cause forgetting: Retrieval dynamics in long-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20* (3), 1063-1087. Retrieved from http://www.apa.org/pubs/journals/xlm/index.aspx

Anderson, M. C., & McColluch, K. C. (1999). Integration as a general boundary condition on retrieval-induced forgetting. *Journal of Experimental Psychology: Learning, Memory and Cognition, 25* (3), 608 – 629. doi:10.1037/0278-7399.25.3.608

Aslan, A., & Bäuml, K. H. (2011). Individual differences in working memory capacity predict retrieval-induced forgetting. *Journal of Experimental Psychology: Learning, Memory and Cognition, 37* (1), 264 – 269. doi: 10.1037/a0021324

Aslan, A., Bäuml, K. H., & Pastötter, B. (2007). No inhibitory deficit in older adults' episodic memory. *Psychological Science, 18* (1), doi:10.1111/j.1467-9280.2007.01851.x

Balota, D. A., Cortese, M. J., Duchek, J. M., Adams, D., Roediger, H. L. III, McDermott, K. B. & Yerys, B. E. (1999). Veridical and false memories in healthy older adults and in dementia of the Alzheimer's type. *Cognitive Neuropsychology, 16* (3-5), 361 – 384. doi:10.1080/026432999380834

Barnier, A. J., Hung, L., & Conway, M. A. (2004). Retrieval-induced forgetting of emotional and

unemotional autobiographical memories. *Cognition & Emotion, 18* (4), 457-477. doi: 10.1080/02699930304000392

Basden, B. H., Basden, D. R., & Morales, E. (2003). The role of retrieval practice in directed forgetting. *Journal of Experimental Psychology: Learning, Memory and Cognition, 29* (3), 389 – 397. doi:10.1037/028-7393.29.3.389

Battig, W. F., & Montague, W. E. (1969). Category norms for verbal items in 56 categories: A replication and extension of the Connecticut norms. *Journal of Experimental Psychology, 80,* 1 – 46. doi:10.1037/h0027577

Bauer, B., & Gourgouvelis, J. E. (2009). Validation of the van Overshelde et al. (2004) category norms: Results from five experiments. *Current psychology letters* [Online], *25* (1). Retrieved from http://cpl.revues.org/index4802.html

Bäuml, K. H., & Kuhbander, C. (2007). Remembering can cause forgetting – but not in negative moods. *Psychological Science, 18,* 111 – 115. doi:10.1111/j.1467-9280.2007.01857.x

Blair, I. V., Lenton, A. P., & Hastie, R. (2002). The reliability of the DRM paradigm as a measure of individual differences in false memories. *Psychological Bulletin & Review, 9* (3), 590 – 596. Retrieved from http://www.psychonomic.org/PBR/forthcoming.htm

Broadbent, D.E., Cooper, P.F., FitzGerald, P., & Parkes, K.R. (1982). The Cognitive Failures Questionnaire (CFQ) and its correlates. *British Journal of Clinical Psychology*, *21*, 1 – 16. Retrieved from http://www.blackwellpublishing.com/journal.asp?ref=0144-6657&site=1

Brownell, W. A. (1933). On the accuracy with which reliability may be measured by correlating test halves. *Journal of Experimental Education, 1,* 204 – 215. Retrieved from http://www.tandf.co.uk/journals/titles/00220973.asp

Butler, K. M., Williams, C. C., Zacks, R. T., & Macki, R. H. (2001). A limit on retrieval-

induced forgetting. *Journal of Experimental Psychology, Learning, Memory & Cognition, 27* (5), 1314 – 1319. Retrieved from http://www.apa.org/pubs/journals/xlm/index.aspx

Cicchetti, D. V. (2001). The precision of reliability and validity estimates re-visited: Distingquishing between clinical and statistical significance of sample size requirements. *Journal of Clinical and Experimental Neuropsychology, 23* (5), 695 – 700. doi:10.1076/jcen.23.5.695.1249

Ciranni, M. A., & Shimamura, A. P. (1999). Retrieval-induced forgetting in episodic memory. *Journal of Experimental Psychology: Learning, Memory & Cognition, 25* (6), 1403 – 1414. doi:10.1037/0278-7393.25.6.1403

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20* (1), 37 – 46. doi:10.1177/001316446002000104

Cohen, L., Manion, L., & Morrison, K. (2007). Research methods in education (6th ed.). New York, NY: Routledge.

Cook, D. A., & Beckman, T. J. (2006). Current concepts in validity and reliability for psychometric instruments: Theory and Application. *The American Journal of Medicine, 119,* 166.e7 – 166.e16. doi:10.1016/j.amjmed.2005.10.036

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78,* 98 – 104. Retrieved from http://www.apa.org/pubs/journals/apl/index.aspx

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. New York, NY: Harcourt Brace Jovanovich College Publishers.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*

(3), 297 334.

Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of Experimental Psychology, 58,* 17-22. doi:10.1037/h0046671

DeVellis, R. F. (2003). Scale development: Theory and applications (2nd ed.). *Applied Social Research Methods Series, 26*.

Dunn, E. W., & Spellman, B. A. (2003). Forgetting by remember: Stereotype inhibition through rehearsal of alternative aspects of identity. *Journal of Experimental Social Psychology, 39* (5), 420-433. doi: 10.1016/S0022-1031(03)00032-5

Field, A. P. (2009). Discovering statistics using SPSS. London, ENG: Sage.

Ford, R. M., Keating, S. & Patel, R. (2002). Retrieval-induced forgetting: A developmental study. *British Journal of Developmental Psychology, 22* (4), 585 – 603. doi:10.1348/026151004237872

Frisbie, D. A. (2005). Measurement 101: Some fundamentals revisited. *Educational Measurement: Issues and Practice, 24* (3), 21 – 28. doi:10.1111/j.1745-3992.2005.00016.x

Gall, M. D., Gall, J. P., & Borg, W. R. (2007). Educational research: An introduction (7th ed.). Boston, MA: Allyn & Bacon.

García-Bajos, E., Migueles, M., & Anderson, M. C. (2008). Script knowledge modulates retrieval-induced forgetting for eyewitness events. *Memory, 17* (1), 92 – 103. doi:10.1080/09658210802572454

Gómez-Ariza, C. J., Lechuga, M. T., Pelegrina, S., & Bejo, M. T. (2005). Retrieval-induced forgetting in recall and recognition of thematically related and unrelated sentences. *Memory & Cognition, 33* (8), 1431-1441. Retrieved from http://mc.psychonomic-

journals.org/

Harnishfeger, K. K., & Bjorklund, D. F. (1994). A developmental perspective on individual differences in inhibition. *Learning and Individual Differences, 6,* 331 – 355. doi:10.1016/1041-6080(94)90021-3

Harnishfeger, K. K., & Pope, R. S. (1996). Intending to forget: The development of cognitive inhibition in directed forgetting. *Journal of Experimental Child Psychology, 62,* 292 – 315. doi:10.1006/jecp.1996.0032

Harris, C. B., Sharman, S. J., Barnier, A. J., & Moulds, M. L. (2010). Mood and retrieval-induced forgetting of positive and negative autobiographical memories. *Applied Cognitive Psychology, 24* (3), 399-413. doi:10.1002/acp.1685

Heffner, C. L. (2004). Research methods for education, psychology and the social sciences. *Allpsych* [Online]. Retrieved from http://www.allpsych.com/researchmethods

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (2003). Applied statistics for the behavioural sciences (5th ed.). Boston, MA: Houghton Mifflin.

Johansson, M., Aslan, A., Bäuml, K. H., Gäbel, A. & Mecklinger, A. (2007). Electrophysical correlates of retrieval-induced forgetting. *Cerebral Cortex, 17* (6), 1335 – 1341. doi: 10.1093/cercor/bhl044

Jones, L. W. (2010). Does increasing metacognitive awareness alleviate retrieval-induced forgetting effects? (Master's dissertation). Retrieved from http://proquest.umi.com/pqdlink?did=2065805801&Fmt=2&clientId=79356&RQT=309&VName=PQD

Kindt, M., Bierman, D., & Brosschot, J. F. (1996). Stroop versus Stroop: Comparison of a card format and a single-trial format of the standard color-word Stroop task and the emotional

Stroop task. *Personality and Individual Differences, 21* (5), 653 – 661. doi:10.1016/0191-8869(96)00133-X

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56* (5), 746 – 759. doi:10.1177/0013164496056005002

Kline, T. J. B. (2005). Classical test theory: Assumptions, equations, limitation, and item analyses. In T. J. B. Kline (Ed.) *Psychological testing: A practical approach to design and evaluation* (pp. 91-105). Thousand Oaks, CA: Sage Publications, Inc.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2,* 151-160. doi:10.1007/BF02288391

Larson, G. E., Alderton, D. L., Neideffer, M., & Underhill, E. (1997). Further evidence on dimensionality and correlates of the Cognitive Failures Questionnaire. *British Journal of Psychology, 88,* 29 – 38. Retrieved from http://onlinelibrary.wiley.com/journal/10.1111/%28ISSN%292044-8295

Lechuga, M. T., Moreno, V., Pelegrina, S., Gómez-Ariza, C. J., & Bajo, M. T., (2006). Age differences in memory control: Evidence from updating and retrieval-practice tasks. *Acta Psychologica, 123,* 279 – 298. doi:10.1016/j.actpsy.2006.01.006

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley Publishing Company.

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin, 109* (2), 163 – 203. doi:10.10.37/0033-2909.109.2.163

MacLeod, M. D. (2002). Retrieval-induced forgetting in eyewitness memory: Forgetting as a consequence of remembering. *Applied Cognitive Psychology, 16* (2), 135 – 149. doi:10.1002/acp.782

MacLeod, M. D., Saunders, J., & Chalmers, L. (2010). Retrieval-induced forgetting: The unintended consequences of unintended forgetting. In G. M. Davies & D. Wright (Eds.), *Current issues in applied memory research*. Hove, UK: Psychology Press.

Macrae, C. N., & MacLeod, M. D. (1999). On recollections lost: When practice makes imperfect. *Journal of Personality and Social Psychology, 77* (3), 463 – 473. Retrieved from http://www.apa.org/pubs/journals/psp/index.aspx

Migueles, M., & García-Bajos, E., (2007). Selective retrieval and induced forgetting in eyewitness memory. *Applied Cognitive Psychology, 21,* 1157 – 1172. doi:10.1002/acp.1323

Marche, T. A., Briere, J. L., & von Baeyer, C. (2011). Individual differences in children's ability to remember and forget negative experiences. Paper presented at the *Society for Research in Child Development* international biennial meeting in Montreal, Canada (March 31 – April 2).

Mather, M., Henkel, L. A., & Johnson, M. K. (1997). Evaluating characteristics of false memories: Remember/know judgements and memory characteristics questionnaire compared. *Memory & Cognition, 25* (6), 826 – 837. dio:10.3758/BF03211327

Matthews, G., Coyle, K., & Craig, A. (1990). Multiple factors of cognitive failure and their relationships with stress vulnerability. *Journal of Psychopathology and Behavioral Assessment, 12,* 49 – 65. doi:10.1007/BF00960453

McDermott , K. B., & Roediger, H. L. (1998). Attempting to avoid illusory memories: Robust false recognition of associates persists under conditions of explicit warnings and immediate testing. *Journal of Memory and Language, 39,* 508 – 520. doi:10.1006/jmla.1998.2582

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.), pp. 13-103. New York, NY: Macmillan.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice, Winter,* 5 - 8. Retrieved from http://www.wiley.com/bw/journal.asp?ref=0731-1745

Moulin, C. J. A., Perfect, T. J., Conway, M. A., North, A. S., Jones, R. W., Niamh, J. (2002). Retrieval-induced forgetting in Alzheimer's disease. *Neuropsychologia, 40* (7), 862-867. Retrieved from http://www.elsevier.com/wps/find/journaldescription.cws_home/247/description#description

Pollina, L. K., Greene, A. L., Tunick, R. H., & Puckett, J. M. (1992). Dimensions of everyday memory in young adulthood. *British Journal of Psychology, 83,* 305 – 321. Retrieved from http://onlinelibrary.wiley.com/journal/10.1111/%28ISSN%292044-8295

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21* 803-814. Retrieved from http://www.apa.org/pubs/journals/xlm/index.aspx

Shaw, J. S., Bjork, R. A., & Handal, A. (1995). Retrieval-induced forgetting in an eyewitness-memory paradigm. *Psychonomic Bulletin & Review, 2* (2), 249 – 253. doi:10.3758/BF03210965

Stöber, J. (2001). The social desirability scale – 17 (SDS-17): Convergent validity, discriminant validity, and relationship with age. *European Journal of Psychological Assessment, 17,* 222-232. Retrieved from http://www.hogrefe.com/periodicals/european-journal-of-

psychological-assessment/

Storm, B. C., Bjork, E. L., & Bjork, R. A. (2005). Social metacognitive judgments: The role of retrieval-induced forgetting in person memory and impressions. *Journal of Memory and Language, 52* (4), 535-550. doi:10.1016/j.jml.2005.01.008

Traub, R. E. (1997). Classical test theory in historical perspective. *Educational Measurement: Issues and Practice, 16,* 8-14. doi:10.1111/j.1745-3992.1997.tb00603.x

Traub, R. E., & Rowley, G. L. (1991). An NCME instructional module on understanding reliability. *Educational Measurement: Issues and Practice, 10* (1), 37 – 45. Retrieved from http://www.wiley.com/bw/journal.asp?ref=0731-1745

van Overschelde, J. P., Rawson, K. A., & Dunlosky, J. (2004). Category norms: An updated and expanded version of the Battig and Montague (1969) norms. *Journal of Memory and Language, 50* (3), 289 – 335. doi:10.1016/j.jml.2003.10.003

Wessel, I. & Hauer, B. (2006). Retrieval-induced forgetting of autobiographical memory details. *Cognition & Emotion, 20* (3-4), 430 – 447. doi:10.1080/02699930500342464

Wilson, S. P., & Kipp, K. (1998). The development of efficient inhibition: Evidence from directed forgetting tasks. *Developmental Review, 18*, 86 – 123. doi:10.1006/drev.1997.0445

APPENDIX A: COUNTERBALANCING ORDERS

The different counterbalancing orders that were used to ensure that order effects did not impact the degree of RIF obtained for each task. *Words A* refers to the first set of word list materials created, and *Words B* refers to the second set of word lists created (see Appendix B for the wordlists). *Sentences* refers to the facts used in sentence format for the RIF task (see Appendix D for the facts).

Order of Materials	Task Order	Phase I	Phase II
1	1	Words A	Words A
	2	Words B	Sentences
2	1	Words A	Words A
	2	Sentences	Words B
3	1	Words B	Words A
	2	Words A	Sentences
4	1	Words B	Words B
	2	Sentences	Words A
5	1	Sentences	Words B
	2	Words A	Words A
6	1	Sentences	Words A
	2	Word A	Words B
7	1	Sentences	Words A
	2	Words B	Words B
8	1	Words B	Words B
	2	Words A	Sentences

Order of Materials	Task Order	Phase I	Phase II
9	1	Words A	Words A
	2	Words B	Sentences
10	1	Words B	Sentences
	2	Words A	Words A
11	1	Words A	Words B
	2	Sentences	Words A
12	1	Words B	Sentences
	2	Words A	Words B
13	1	Words A	Words B
	2	Words B	Sentences
14	1	Words A	Sentences
	2	Words B	Words B
15	1	Words B	Words A
	2	Sentences	Words B
16	1	Sentences	Words B
	2	Words B	Words A

APPENDIX B: WORD LISTS AND RETRIEVAL-PRACTICE FRAGMENTS

Categories, exemplars and exemplar retrieval-practice fragments for each set of word list materials.

Set	Category	Items	Retrieval-Practice Fragments
Word List A	Flower	Orchid	Or_____
		Petunia	Pe_____
		Lilac	Li_____
		Carnation	Ca_____
		Daffodil	Da_____
		Tulip	Tu_____
	Bird	Pigeon	Pi_____
		Bluejay	Bl_____
		Robin	Ro_____
		Eagle	Ea_____
		Cardinal	Ca_____
		Seagull	Se_____
	Footwear	Slippers	Sl_____
		Cleats	Cl_____
		Sneakers	Sn_____
		Heels	He_____
		Loafers	Lo_____
		Sandals	Sa_____
	Car	Nissan	Ni_____
		Ford	Fo_____
		Toyota	To_____
		Dodge	Do_____
		Mustang	Mu_____
		Cadillac	Ca_____
	Drug	Crack	Cr_____
		Marijuana	Ma_____
		Heroin	He_____
		Ecstasy	Ec_____
		Advil	Ad_____
		Cocaine	Co_____
	Sport	Volleyball	Vo_____
		Swimming	Sw_____
		Football	Fo_____
		Tennis	Te_____
		Lacrosse	La_____
		Hockey	Ho_____

Set	Category	Items	Retrieval-Practice Fragments
Word List A	Tree	Maple	Ma_____
		Aspen	As_____
		Spruce	Sp_____
		Birch	Bi_____
		Redwood	Re_____
		Pine	Pi_____
	Tool	Wrench	Wr_____
		Drill	Dr_____
		Screwdriver	Sc_____
		Hammer	Ha_____
		Nail	Na_____
		Level	Le_____
	Vegetable	Lettuce	Le_____
		Broccoli	Br_____
		Corn	Co_____
		Potato	Po_____
		Onion	On_____
		Spinach	Sp_____
	Dance	Tango	Ta_____
		Jazz	Ja_____
		Waltz	Wa_____
		Swing	Sw_____
		Ballroom	Ba_____
		Modern	Mo_____

Set	Category	Items	Retrieval-Practice Fragments
Word List B	Liquid	Juice	Ju_____
		Soda	So_____
		Beer	Be_____
		Blood	Bl_____
		Mercury	Me_____
		Gasoline	Ga_____
	Instrument	Flute	Fl_____
		Piano	Pi_____
		Clarinet	Cl_____
		Saxophone	Sa_____
		Cello	Ce_____
		Viola	Vi_____

Set	Category	Items	Retrieval-Practice Fragments
Word List B	Clothing	Pants	Pa_____
		Underwear	Un_____
		Shoes	Sh_____
		Hats	Ha_____
		Jacket	Ja_____
		Sweater	Sw_____
	Military	General	Ge_____
		Captain	Ca_____
		Private	Pr_____
		Colonel	Co_____
		Major	Ma_____
		Officer	Of_____
	Fish	Salmon	Sa_____
		Trout	Tr_____
		Catfish	Ca_____
		Shark	Sh_____
		Dolphin	Do_____
		Blowfish	Bl_____
	Insect	Mosquito	Mo_____
		Beetle	Be_____
		Grasshopper	Gr_____
		Butterfly	Bu_____
		Roach	Ro_____
		Gnats	Gn_____
	Fabric	Polyester	Po_____
		Wool	Wo_____
		Nylon	Ny_____
		Satin	Sa_____
		Denim	De_____
		Rayon	Ra_____
	Crime	Murder	Mu_____
		Rape	Ra_____
		Stealing	St_____
		Assault	As_____
		Burglary	Bu_____
		Arson	Ar_____

Set	Category	Items	Retrieval-Practice Fragments
Word List B	Wood	Table	Ta_____
		Desk	De_____
		House	Ho_____
		Floor	Fl_____
		Dresser	Dr_____
		Cabinet	Ca_____
	Fruit	Strawberry	St_____
		Orange	Or_____
		Pear	Pe_____
		Grapes	Gr_____
		Pineapple	Pi_____
		Banana	Ba_____

APPENDIX C: ITEM DEVELOPMENT SHEET

The following is a sample of the item development sheets provided to judges to ensure that study items belonged to only one category.

RELIABILITY OF RIF ITEM DEVELOPMENT ID:_____

SET A MATERIALS

INSTRUCTIONS: The following table provides a list of items along with 10 different categories. Please read each item listed on the left hand side of the table and circle the category name that the item best falls under (e.g., "rat" is a type of "animal," thus it would fit under an "animal" category). If the item appears to belong to more than one category, please circle all relevant category names.

ITEM		CATEGORY									
1	Football	Flower	Drug	Sport	Bird	Dance	Vegetable	Fabric	Tool	Footwear	Car
2	Wool	Flower	Drug	Sport	Bird	Dance	Vegetable	Fabric	Tool	Footwear	Car
3	Hammer(s)	Flower	Drug	Sport	Bird	Dance	Vegetable	Fabric	Tool	Footwear	Car
4	Drill	Flower	Drug	Sport	Bird	Dance	Vegetable	Fabric	Tool	Footwear	Car
5	Loafer(s)	Flower	Drug	Sport	Bird	Dance	Vegetable	Fabric	Tool	Footwear	Car
6	Bluejay	Flower	Drug	Sport	Bird	Dance	Vegetable	Fabric	Tool	Footwear	Car
7	Cardinal	Flower	Drug	Sport	Bird	Dance	Vegetable	Fabric	Tool	Footwear	Car
8	Corn	Flower	Drug	Sport	Bird	Dance	Vegetable	Fabric	Tool	Footwear	Car
9	Robin(s)	Flower	Drug	Sport	Bird	Dance	Vegetable	Fabric	Tool	Footwear	Car
10	Tulip	Flower	Drug	Sport	Bird	Dance	Vegetable	Fabric	Tool	Footwear	Car
11	Clog(s)	Flower	Drug	Sport	Bird	Dance	Vegetable	Fabric	Tool	Footwear	Car
12	Ecstasy	Flower	Drug	Sport	Bird	Dance	Vegetable	Fabric	Tool	Footwear	Car
13	Lacrosse	Flower	Drug	Sport	Bird	Dance	Vegetable	Fabric	Tool	Footwear	Car
14	Advil	Flower	Drug	Sport	Bird	Dance	Vegetable	Fabric	Tool	Footwear	Car
15	Nylon	Flower	Drug	Sport	Bird	Dance	Vegetable	Fabric	Tool	Footwear	Car
16	Lily	Flower	Drug	Sport	Bird	Dance	Vegetable	Fabric	Tool	Footwear	Car
17	Petunia(s)	Flower	Drug	Sport	Bird	Dance	Vegetable	Fabric	Tool	Footwear	Car
18	Crack	Flower	Drug	Sport	Bird	Dance	Vegetable	Fabric	Tool	Footwear	Car
19	Ford	Flower	Drug	Sport	Bird	Dance	Vegetable	Fabric	Tool	Footwear	Car
20	Pigeon(s)	Flower	Drug	Sport	Bird	Dance	Vegetable	Fabric	Tool	Footwear	Car
21	Level	Flower	Drug	Sport	Bird	Dance	Vegetable	Fabric	Tool	Footwear	Car

22	Heroin	Flower	Drug	Sport	Bird	Dance	Vegetable	Fabric	Tool	Footwear	Car
23	Marijuana	Flower	Drug	Sport	Bird	Dance	Vegetable	Fabric	Tool	Footwear	Car
24	Cocaine	Flower	Drug	Sport	Bird	Dance	Vegetable	Fabric	Tool	Footwear	Car
25	Dodge	Flower	Drug	Sport	Bird	Dance	Vegetable	Fabric	Tool	Footwear	Car
26	Volleyball	Flower	Drug	Sport	Bird	Dance	Vegetable	Fabric	Tool	Footwear	Car
27	Slipper(s)	Flower	Drug	Sport	Bird	Dance	Vegetable	Fabric	Tool	Footwear	Car
28	Nail(s)	Flower	Drug	Sport	Bird	Dance	Vegetable	Fabric	Tool	Footwear	Car
29	Sneaker(s)	Flower	Drug	Sport	Bird	Dance	Vegetable	Fabric	Tool	Footwear	Car
30	Eagle	Flower	Drug	Sport	Bird	Dance	Vegetable	Fabric	Tool	Footwear	Car
31	Carnation(s)	Flower	Drug	Sport	Bird	Dance	Vegetable	Fabric	Tool	Footwear	Car
32	Wrench	Flower	Drug	Sport	Bird	Dance	Vegetable	Fabric	Tool	Footwear	Car
33	Orchid(s)	Flower	Drug	Sport	Bird	Dance	Vegetable	Fabric	Tool	Footwear	Car
34	Onion(s)	Flower	Drug	Sport	Bird	Dance	Vegetable	Fabric	Tool	Footwear	Car
35	Sandal(s)	Flower	Drug	Sport	Bird	Dance	Vegetable	Fabric	Tool	Footwear	Car
36	Waltz	Flower	Drug	Sport	Bird	Dance	Vegetable	Fabric	Tool	Footwear	Car
37	Jazz	Flower	Drug	Sport	Bird	Dance	Vegetable	Fabric	Tool	Footwear	Car
38	Swimming	Flower	Drug	Sport	Bird	Dance	Vegetable	Fabric	Tool	Footwear	Car
39	Swing	Flower	Drug	Sport	Bird	Dance	Vegetable	Fabric	Tool	Footwear	Car
40	Cadillac	Flower	Drug	Sport	Bird	Dance	Vegetable	Fabric	Tool	Footwear	Car
41	Toyota	Flower	Drug	Sport	Bird	Dance	Vegetable	Fabric	Tool	Footwear	Car
42	Lilac	Flower	Drug	Sport	Bird	Dance	Vegetable	Fabric	Tool	Footwear	Car
43	Nissan	Flower	Drug	Sport	Bird	Dance	Vegetable	Fabric	Tool	Footwear	Car
44	Hockey	Flower	Drug	Sport	Bird	Dance	Vegetable	Fabric	Tool	Footwear	Car
45	Denim	Flower	Drug	Sport	Bird	Dance	Vegetable	Fabric	Tool	Footwear	Car
46	Spinach	Flower	Drug	Sport	Bird	Dance	Vegetable	Fabric	Tool	Footwear	Car
47	Rayon	Flower	Drug	Sport	Bird	Dance	Vegetable	Fabric	Tool	Footwear	Car
48	Potato	Flower	Drug	Sport	Bird	Dance	Vegetable	Fabric	Tool	Footwear	Car
49	Tango	Flower	Drug	Sport	Bird	Dance	Vegetable	Fabric	Tool	Footwear	Car
50	Ballroom	Flower	Drug	Sport	Bird	Dance	Vegetable	Fabric	Tool	Footwear	Car
51	Heels	Flower	Drug	Sport	Bird	Dance	Vegetable	Fabric	Tool	Footwear	Car
52	Polyester(s)	Flower	Drug	Sport	Bird	Dance	Vegetable	Fabric	Tool	Footwear	Car

53	Seagull(s)	Flower	Drug	Sport	Bird	Dance	Vegetable	Fabric	Tool	Footwear	Car
54	Broccoli	Flower	Drug	Sport	Bird	Dance	Vegetable	Fabric	Tool	Footwear	Car
55	Mustang	Flower	Drug	Sport	Bird	Dance	Vegetable	Fabric	Tool	Footwear	Car
56	Modern	Flower	Drug	Sport	Bird	Dance	Vegetable	Fabric	Tool	Footwear	Car
57	Lettuce	Flower	Drug	Sport	Bird	Dance	Vegetable	Fabric	Tool	Footwear	Car
58	Satin	Flower	Drug	Sport	Bird	Dance	Vegetable	Fabric	Tool	Footwear	Car
59	Tennis	Flower	Drug	Sport	Bird	Dance	Vegetable	Fabric	Tool	Footwear	Car
60	Screwdriver (s)	Flower	Drug	Sport	Bird	Dance	Vegetable	Fabric	Tool	Footwear	Car

Please answer the following demographic questions.

1. What is your age (in years)?:_____
2. What is your gender (circle one)?: Male / Female

If you have any other comments or suggestions, please feel free to include them by the words or here:

_____

_____

_____

_____

Thank you for your help!

APPENDIX D: FACTS LIST AND RETRIEVAL-PRACTICE FRAGMENTS

(Macrae & MacLeod, 1999)

Island names, associated facts and retrieval-practice fragments.

Island	Items	Retrieval-Practice Fragments
Bilu	The main cash crop in Bilu is cocoa.	The main cash crop in Bilu is co_____.
	Bilu was first settled by sailors from Spain.	Bilu was first settled by sailors from Sp_____.
	Most visitors to Bilu come from Chile.	Most visitors to Bilu come from Ch_____.
	All farmers in Bilu keep goats.	All farmers in Bilu keep go_____.
	Bilu's only major export is copper.	Bilu's only major export is co_____.
	In Bilu, 67% of the population own cars.	In Bilu, 67% of the population own ca_____.
	In Bilu, most people gamble.	In Bilu, most people ga_____.
	The wettest month in Bilu is March.	The wettest month in Bilu is Ma_____.
	There are 260 varieties of spider in Bilu	There are 260 varieties of sp_____.
	The shops in Bilu close on Tuesdays.	The shops in Bilu close on Tu_____.
Tok	Tok's national day is in June.	Tok's national day is in Ju_____.
	The staple food of Tok is maize.	The staple food of Tok is ma_____.
	There are no horses on Tok.	There are no ho_____ on Tok.
	42% of Tok's population own their own homes.	42% of Tok's population own their own ho_____.
	93% of people on Tok own a bicycle.	93% of people on Tok own a bi_____.
	Fishing is the national past-time on Tok.	Fishing is the national past-time on Tok.
	Houses on Tok are traditionally made of wood.	Houses on Tok are traditionally made of wo_____.
	Tok's nearest neighbour is Tiawan.	Tok's nearest neighbour is Ti_____.
	French is the national language of Tok.	Fr_____ is the national language of Tok.
	Tok's currency is the dollar.	Tok's currency is the do_____.

APPENDIX E: THE COGNITIVE FAILURES QUESTIONNAIRE

(Broadbent, Cooper, Fitzgerald & Parkes, 1982)

The following questions are about minor mistakes which everyone makes from time to time, but some of which happen more often than others. We want to know how often these things have happened to you in the past 6 months. Please circle the appropriate number.

		Very often	Quite often	Occasion-ally	Very rarely	Never
1.	Do you read something and find you haven't been thinking about it and must read it again?	4	3	2	1	0
2.	Do you find you forget why you went from one part of the house to the other?	4	3	2	1	0
3.	Do you fail to notice signposts on the road?	4	3	2	1	0
4.	Do you find you confuse right and left when giving directions?	4	3	2	1	0
5.	Do you bump into people?	4	3	2	1	0
6.	Do you find you forget whether you've turned off a light or a fire or locked the door?	4	3	2	1	0
7.	Do you fail to listen to people's names when you are meeting them?	4	3	2	1	0
8.	Do you say something and realize afterwards that it might be taken as insulting?	4	3	2	1	0
9.	Do you fail to hear people speaking to you when you are doing something else?	4	3	2	1	0
10.	Do you lose your temper and regret it?	4	3	2	1	0
11.	Do you leave important letters unanswered for days?	4	3	2	1	0

		Very often	Quite often	Occasion-ally	Very rarely	Never
12.	Do you find you forget which way to turn on a road you know well but rarely use?	4	3	2	1	0
13.	Do you fail to see what you want in a supermarket (although it's there)?	4	3	2	1	0
14.	Do you find yourself suddenly wondering whether you've used a word correctly?	4	3	2	1	0
15.	Do you have trouble making up your mind?	4	3	2	1	0
16.	Do you find you forget appointments?	4	3	2	1	0
17.	Do you forget where you put something like a newspaper or a book?	4	3	2	1	0
18.	Do you find you accidentally throw away the thing you want and keep what you meant to throw away – as in the example of throwing away the matchbox and putting the used match in your pocket?	4	3	2	1	0
19.	Do you daydream when you ought to be listening to something?	4	3	2	1	0
20.	Do you find you forget people's names?	4	3	2	1	0
21.	Do you start doing one thing at home and get distracted into doing something else (unintentionally)?	4	3	2	1	0
22.	Do you find you can't quite remember something although it's "on the tip of your tongue"?	4	3	2	1	0

23.	Do you find you forget what you came to the shops to buy?	4	3	2	1	0
24.	Do you drop things?	4	3	2	1	0
25.	Do you find you can't think of anything to say?	4	3	2	1	0

APPENDIX F: SOCIAL DESIRABILITY SCALE – 17

(Stöber, 2001)

Instructions: Below you will find a list of statements. Please read each statement carefully and decide if that statement describes you or not. If it describes you, check the word "true"; if not, check the word "false."

1. I sometimes litter.	True	False
2. I always admit my mistakes openly and face the potential negative consequences.	True	False
3. In traffic I am always polite and considerate of others.	True	False
4. I have tried illegal drugs (for example, marijuana, cocaine, etc.).	True	False
5. I always accept others' opinions, even when they don't agree with my own.	True	False
6. I take out my bad moods on others now and then.	True	False
7. There has been an occasion when I took advantage of someone else.	True	False
8. In conversations I always listen attentively and let others finish their sentences.	True	False
9. I never hesitate to help someone in case of emergency.	True	False
10. When I have made a promise, I keep it – no ifs, ands or buts.	True	False
11. I occasionally speak badly of others behind their back.	True	False
12. I would never live off other people.	True	False
13. I always stay friendly and courteous with other people, even when I am stressed out.	True	False
14. During arguments I always stay objective and matter of fact.	True	False
15. There has been at least one occasion when I failed to return an item that I borrowed.	True	False
16. I always eat a healthy diet.	True	False
17. Sometimes I only help because I expect something in return.	True	False

APPENDIX G: DEMOGRAPHICS

To help describe the participants who took part in the study, please answer the following questions.

1. What is your age (in years)? _____

2. What is your gender (circle one)? Male / Female

3. What is your first language? _____

APPENDIX H: VISUAL SEARCH TASK

On both sides of the page are random capital letters. Your task is to stroke out (put a line through the letter with your pencil) as many vowels as you can find. Vowels include the letters: A, E, I, O, U, and Y.

A S I J T K M D Y P A J R B I F X Q T B N O P G S G T B W Q C G J J L M T Q P I Y N G D

I I K L M B T F J O H O M B T I T E Y P X X C Y W Q P Z M W O N X C I V U E V R B G T

Y H J U N A S D F T G H B V E Y J K O P Y W X B G F H Y T K L M N Y D U J G D B M P

U S G J L P I Y R W Z C B M N V X A D G J L O U T E Q W G F C J M N O L H F U K M V

X S W D Q T P L B C X J I G N M K F E D S Q D C P J H Q L D K W O N X H F U H G I U

Q P S D F J A W E W E R J S A D O F J W E R W O R G J O E R Y T Q W P Q Y E T B B M

A P W I R D G D J J P Q Q S L F H O Q O F S F S H O H W H W R J F O V Y T Q R I O U S

F H K J U B Y C T Z E B U H N O P M K I N J U V G C T S E F K O A S Y W O H E F Q F P

H G W O U F C N A O J Y A R P T O U I B J N A V Y S M I E B D S O R H C Y P M C U S

W D J W F D U W H C G U H D B I D Q H E F H O G S V H O Q P O B N V E O R V O N T

K K C L H F B T F O U W H L Q N U I N S S N O V S U Y S N V E U E A N U V L P M W E

W N E R T U Y C I N A H E F G I U E H F O S L B G I U W H F D O Q B W G O W N F P W

G F J W O P W E R T N V Q Y E R N B V Q W U T Y H F P O Q G H W P O R H T N U I E

G H V N I O U E R H G N A P P A W E P W E R G H E R I J G H P A W V W P S D G H N A

F S K C P O I G H E R F I Y U O I U Y T R T S W D F G H K L N M B V C X S U Y I T F U

H G V C F J Y F U K G L H J B K I O I U Y U T W R A S H D F L G I V K B H I O I U Y T R

W S D G F H K G J K V N M B C G S D F K G J H V B Q E R S D F G C V B R T D F G J H

V M B N O K J H N P O K J N H G B A E R S T D Y F U I O L K J H G F D S F C G V B H J

N M M N B V C E R S D F G X U Y F J H V N P O I K J H N O H V M P O I J H E S D A O

W R G V H J T E G U I C D K A F W Q E R O W I E Y T N C W E F P W E L K D J F A W S

E D R F V T G B H N K O I J U H D K S D A O W R G V H J T E G U I J Y T R E S D F T Y

G U H I J L K J H G F D S A Z X C V B N J K I U Y T R D F C V B N J K I U Y T R E W E A

S D F G H V B N J K L O P I U Y T R E W A S D R F T Y U H J B N H J G U Y T F G E W S

G D F X U T Y F J G H V M P O I J H E S D A O W R G V H J T E G U I C B L E R W S U G

F O D H J L S B E T S H U D F G P A I E W Q P Y U W T Z B V K U R W G I E F L A F G I

W A R P H G E A P U W E I R P E I M V I T V T S I R O H J A E O R I H G A O W U E T Y

Q W P E T G H O S N V G J S F D G H A S D C M N V A F I J G H A P I S O F J I W A P Q

A W S E D X R C F V G B H N J M K P O I U Y T R E A S Z D X C F G V H B J N K J M H
N F G D X S Z E R T Y U H J K N H B G V F D E R T Y U H I J K S H G I U A W E N H A V
I U E W N H O P A E W J F P A E O R H G P W A E F G A E S U Y H F G I D H G I A S G C
N A P W E Y T U W G F C J M N O L H F U K M V X S W D Q T P L B C X J I G N M K F E
D S Q D C P J H Q L D K W O E Y N Q C H F I U B V V N O W H R U Y W B R N E O R U
Y P I A W E T B F E I R H G E I O N R U A V H M P O E A S H X O I N A E R H G A N O E
R H C P A D O F J W E R W O R G J O E R Y T Q W P Q Y E T B B M A P W I R D G D J J
P Q Q S L F H O Q B W G O W N F P W G F U K G L H J B K I O I U Y U T W R A S H D F
L G I V K B H I O I U Y T R W S D G F H K G J K V N M B C G S D F K G J W O P W E R
T N V Q Y E R N B V Q W U T Y H F P A S H D F L G I V K B H I O I U Y T R W S D G F
H K G J K V N M B C G S D F K G J H V B Q E R S D F G C V B R T D F G J H V M B N O
K J H N P O K J N H G B A E R S O Q G H W B N H J G U Y T F G E W S G D F X U T Y F
J G H V M P O I J H E S D A O W R G V H J T E G U I C D K A F W Q E R O W I E Y T N C
W E F P W E L K D J F A W S E D R F V T G B H N K O I J U H D K S D A O W R G V H J
T E G U I C B L E R W S U G F O D H J L S B E T S H U D F G P A I E W Q P J S K W O D
U G H V B N C M X U F V B N J I U Y T R E S X C V B N M K P O I R U Y E U R Y G T B
V A E O R I H G A O W U E T Y Q W P E T G H O S N V G J S F D G H A S D C M N V A F
I J G H A P I S O F J I W A P Q A W S E D X R C F V G B H N J M K P O I U Y T R E A S Z
D X C F G V H B J N U G F B U A V E F O B I A E W U R O Q N C W E H O I E R Y G A B
V O I F H I O B A U W R V P B W A B T Y E P U Y F A U E W I B F V P W E U I F H B E V
A I F H A I P N S F P W O V A S D G R I E U T V N T R H R W G I E F L A F G I W A R P H
G E A P U W E I R P E I M V I T V T S I R O H J A E O R I H G A O W U E T Y Q W P E T
G H O S N V G J S F D G H A S D C M N V A F I J G H A P I S O F J I W A P Q A W S E D
X R C F V G B H N J M K P O I U Y T R E A S Z D X C F G V H B J N K J M H N F G D X S
Z E R T Y U H J K N H B G V F D E R T Y U H I J K S H G I U A W E N H A V I U E W N H
O P A E W J F P A E O R H G P W A E F G A E S U Y H F G I D H G I A S G C N A P W E Y
T U W E N A P W H G A P E R A G H A E I F H G W H U E G F O W E G F B O A V W I H
E F A N P D K W O N X H F U H G I U Q P S D F J A W E W E R J S A D O F J W E R W O
R G J O E R Y T Q W P Q Y E T B B M A P W I R D G D J J P Q Q S L F H O Q O F S F S H
O H W H W R J F O V Y T Q R I O U S F H K J U B Y C T Z E B U H N O P M K I N J U V
G C T S E F K O A S Y W O H W J H E F I A W H O A F S O G J O D J H G O E H G O S N

APPENDIX I: PERSONAL CODE INSTRUCTIONS

In order to link your data across your two testing dates without identifying you personally, please create your personal code by writing down the ***first three letters of your Mother's first name,*** along with the ***day and month of your birth*** (e.g., my Mother's name is Sandra and my birthday is January 25th; my code would be SAN2501).

Please write your code here:

_____	_____	_____
First 3 Letters of Mother's Name	Your Birth Day	Your Birth Month

Date of Phase I:_____

Date of Phase II:_____

Date of Phase III:_____

APPENDIX J: CONSENT FORM

UNIVERSITY OF SASKATCHEWAN

Memory for Related Words and Sentences

(Beh-REB#: 02-381)

You are invited to participate in a research project entitled *Memory for Related Words and Sentences*. Please read this form carefully, and feel free to ask questions you might have.

Student Researcher
Jennifer Briere
Department of Psychology
133B St. Thomas More College
University of Saskatchewan
Phone: 966-8314
Email: jennifer.briere@usask.ca

Research Supervisor
Dr. Tammy Marche
Department of Psychology
440 St. Thomas More College
University of Saskatchewan
Phone: 966-8076
Email: tmarche@stmcollege.ca

Purpose and Procedure: The data collected during this study will be used as part of a research paper. Data collected will be reported in aggregate form to ensure confidentiality. The purpose of this experiment is to determine whether or not individuals' ability to remember the same, or different types of information is a stable characteristic or one that changes (i.e., a stable individual difference). To determine whether or not individual's ability to remember different types of information is similar across time and materials, participants in the study will be asked to sign up for two different 30 minute testing sessions, two weeks apart (total of 1 hour).

If you choose to participate, you will be randomly assigned to one of two groups. The first group will complete two memory tasks at the first testing session using words and sentences, and only one memory task using words at the second testing session. The other group will complete one memory task using words at the first testing session and two memory tasks using words and sentences at the second testing session.

For the memory task using words, you will be asked to study a number of word pairs consisting of a category name and the paired item (e.g., "literature – poem"). After all word pairs have been studied once, you will be given three more opportunities to practice recalling some of the word pairs. Next, you will be asked to complete 5 minutes of simple mathematics and finally, recall as many of the word pairs as you can without guessing.

The second memory task that you will be asked to complete is quite similar to the first memory task, however sentences will be used instead of word pairs. You will first be asked to study a number of sentences that contain facts about two fictitious places. After all of the sentences have been studied once, you will be given three additional opportunities to study some of the sentences. Following 5 minutes of mathematics, you will be asked to recall as many of the sentences as you can without guessing.

Finally, you will be asked to complete a few demographic questions to help describe the participants who completed the study.

Potential Benefits & Risks: There is no guarantee that you will personally benefit from your involvement with the research. There are no known risks associated with participating in the study.

Storage of Data: Data and consent forms will be stored securely at the University of Saskatchewan by the research supervisor for a minimum of five years.

Confidentiality: Please do not put your name or any identifying information (e.g., student number) on your response sheets. In order to link the data across testing sessions, you will be asked to create a unique participant code that will not identify you individually. To create this code, you will be asked to provide the first three letters of your Mother's first name along with the day and month of your birth (e.g., Sandra, Jan. 25th = SAN2501). Your data will be kept completely confidential. Information that is shared will be held in strict confidence and discussed only with the research team. After you have completed the study, your answers will be put in a sealed envelope so that answers cannot be associated with individuals.

Right to Withdraw: Your participation is voluntary, and you can answer only those questions that you are comfortable with. The information that is shared will be held in strict confidence and discussed only with the research team. You may withdraw from the research project for any reason, at any time, without penalty of any sort (e.g., participant credit). If you withdraw from the research project at any time, any data that you have contributed will be destroyed beyond recovery at your request.

Questions: If you have any questions concerning the research project, please feel free to ask at any point; you are also free to contact the researchers at the numbers provided if you have other questions. This research project has been approved on ethical grounds by the University of Saskatchewan Behavioural Research Ethics Board on June 1st, 2010. Any questions regarding your rights as a participant may be addressed to that committee through the Ethics Office (966-2084). Out of town participants may call collect.

Debriefing: Once you have finished the study, you will be provided with a debriefing form that describes the purposes of the study in more detail. If you would like to request a copy of the final results, please contact the researchers at the numbers provided.

Consent to Participate: I have read and understood the description provided; I have had an opportunity to ask questions and my/our questions have been answered. I consent to participate in the research project, understanding that I may withdraw my consent at any time. A copy of this Consent Form has been given to me for my records.

_____	_____
(Name of Participant)	(Date)
_____	_____
(Signature of Participant)	(Signature of Researcher)

APPENDIX K: WORD LIST RETRIEVAL-INDUCED FORGETTING INSTRUCTIONS

Initial Study (Learning) Instructions

(1) Memory for Wordlists: Instructions

The researcher will provide you with a study booklet. Please keep the booklet facedown in front of you until the researcher tells you to begin.

Once instructed to do so, your task will be to study a number of "category – word" pairs, one at a time (e.g., Literature – Poems, Literature – Narrative) for five seconds each. One word pair will be written on each page of the study booklet. Once five seconds have passed, you will hear a tone indicating that you should turn to the next page of your booklet. During the five seconds of study time, do your best to study the pair to remember it as best you can. Once you have studied all words, further instructions will be provided.

Retrieval-Practice Instructions

(2) Memory for Wordlists: Instructions

The researcher will provide you with a practice booklet. Please keep the booklet facedown in front of you until the researcher tells you to begin.

You will now have an opportunity to practice remembering some of the word pairs you just studied. On each page of the practice-booklet will be a fill-in-the-blanks like task for you to complete. The category name and part of the paired word will be provided (e.g., Literature – Po_____). Your task will be to fill-in-the-blank by completing the word fragment with one of the words you just studied (e.g., Literature – Poems). Some fragments will be presented more than once. You can move on to the next page once you have finished filling in each blank, but please do not skip fragments or guess. Once you have completed all the fragments further instructions will be provided.

Distractor Task Instructions

(3) Memory for Sentences: Instructions

The researcher will provide you with two forms. Please keep them face down in front of you until the researcher tells you to begin.

Once the researcher tells you to, please turn over the first form and read the instructions. You will have a total of 5 minutes to complete the forms. If you complete the questions before 5 minutes have passed, please turn over the second form, read the instructions and begin completing it. The researcher will indicate when 5 minutes have passed and further instructions will be provided.

Recall Instructions

(4) Memory for Wordlists: Instructions

The researcher will provide you with a test booklet. Please keep the booklet facedown in front of you until the researcher tells you to begin.

You will now be asked to write down all of the words that you remember studying during the initial study task. At the top of each page of the booklet will be one of the categories from the "category – word" pairs that you studied. After reading the category, please write down as many of the words that you remember studying that fit with that category. Spelling does not matter and order does not matter, just please do not guess. Once you have tried your best to recall as many words as you can that fit with that category, turn the page and move on to the next category. Once you have finished the test booklet, please turn it over and the researcher will provide you with further instructions.

APPENDIX L: FACTS RETRIEVAL-INDUCED FORGETTING INSTRUCTIONS

Initial Study (Learning) Instructions

(1) Memory for Sentences: Instructions

The researcher will provide you with a study booklet. Please keep the booklet facedown in front of you until the researcher tells you to begin.

Once instructed to do so, your task will be to study a number of facts about two different islands, one at a time (e.g., Carten – The national flower of Carten is the lily) for five seconds each. One fact will be written on each page of the study booklet. Once five seconds have passed, you will hear a tone indicating that you should turn to the next page of your booklet. During the five seconds of study time, do your best to study the fact to remember it as best you can. Once you have studied all of the facts, further instructions will be provided.

Retrieval-Practice Instructions

(2) Memory for Sentences: Instructions

The researcher will provide you with a practice booklet. Please keep the booklet facedown in front of you until the researcher tells you to begin.

You will now have an opportunity to practice remembering some of the facts you just studied. On each page of the practice-booklet will be a fill-in-the-blanks like task for you to complete. The island name and part of the fact will be provided (e.g., Carten – The national flower of Carten is the li___). Your task will be to fill-in-the-blank by completing the word fragment with one of the facts that you just studied (e.g., Carten – The national flower of Carten is the lily). Some fragments will be presented more than once. You can move on to the next page once you have finished filling in each blank, but please do not skip fragments or guess. Once you have completed all the fragments further instructions will be provided.

(3) Memory for Sentences: Instructions

The researcher will provide you with two forms. Please keep them face down in front of you until the researcher tells you to begin.

Once the researcher tells you to, please turn over the first form and read the instructions. You will have a total of 5 minutes to complete the forms. If you complete the questions before 5 minutes have passed, please turn over the second form, read the instructions and begin completing it. The researcher will indicate when 5 minutes have passed and further instructions will be provided.

Recall Instructions

(4) Memory for Sentences: Instructions

The researcher will provide you with a test booklet. Please keep the booklet facedown in front of you until the researcher tells you to begin.

You will now be asked to write down all of the sentences that you remember studying during the initial study task. At the top of each page of the booklet will be the name of one of the islands that you studied. After reading the island name, please write down as many of the sentences that you remember studying that fit with that island. Spelling does not matter and order does not matter, just please do not guess. Once you have tried your best to recall as many sentences as you can that fit with that island, turn the page and move on to the next island. Once you have finished the test booklet, please turn it over and the researcher will provide you with further instructions.

APPENDIX M: DEBRIEFING FORM

UNIVERSITY OF SASKATCHEWAN

Memory for Related Words and Sentences

(Beh-REB#: 02-381)

DEBRIEFING FORM

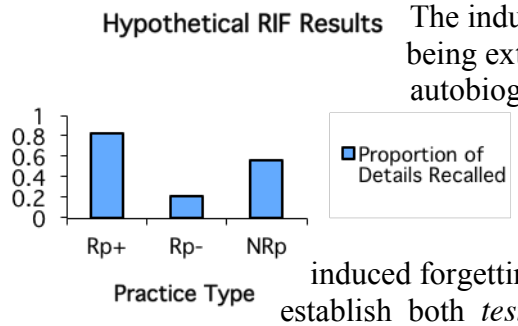
Thank you for your participation in a research project entitled *Memory for Related Words and Sentences*. We really appreciate your help!

Student Researcher
Jennifer Briere
Department of Psychology
133B St. Thomas More College
University of Saskatchewan
Phone: 966-8314
Email: jennifer.briere@usask.ca

Research Supervisor
Dr. Tammy Marche
Department of Psychology
440 St. Thomas More College
University of Saskatchewan
Phone: 966-8076
Email: tmarche@stmcollege.ca

The memory tasks that you completed followed an unintentional forgetting procedure called *Retrieval Induced Forgetting* (RIF; Anderson, Bjork & Bjork, 1994). In this procedure, participants study a number of details (e.g., words, sentences) that relate to different semantic categories. During the retrieval-practice phase, participants practice retrieving or recalling *some* (e.g., half) of the details from one of the categories across three different trials. Following a brief filler task, participants are asked to recall as many of the details as they can remember from the initial study presentation. This procedure results in three different levels of practice: details that received retrieval-practice (Rp+), details that do not receive any retrieval-practice but are from the retrieval-practiced category (Rp-), and the other category that receives no retrieval manipulation (NRp).

Recall data from participants who complete a RIF task typically shows increased recall of the details that received extra practice (Rp+) when compared to the no retrieval-practice baseline (NRp; i.e., a practice effect). However, recall of the details that did not receive any retrieval-practice but were from the practiced category (Rp-) typically show inhibition in memory, resulting in significantly lower recall of Rp- details when compared to the NRp baseline (see the figure below for hypothetical results).



The induced forgetting that results from the RIF procedure is being extended beyond word lists and sentences to autobiographical memory (e.g., Barnier, Hung & Conway, 2001) and is being investigated as a potential memory based intervention. Regardless of such research attention, there is a lack of research establishing the psychometric properties of RIF (e.g., reliability of the induced forgetting obtained). Thus, the goal of the current research is to establish both *test-retest reliability* (i.e., the stability of scores across

testing intervals) and *alternate forms reliability* (i.e., the strength of the relationship between scores obtained on two different forms of the same test) of the RIF procedure. Test-retest reliability estimates will be examined across a two-week delay using the three memory tasks using words. Alternate forms reliability estimates will be established through correlations between the words and sentence memory tasks. It is expected that strong positive correlations will be obtained across all memory tasks indicating that the degree of forgetting obtained when using similar or different materials is quite stable in individuals when the same RIF procedure is used.

If you have any questions, please feel free to contact the researchers at the numbers above. Thanks again for your help with the project!

References

Anderson, M. C., Bjork, R. A., & Bjork, E. L. (1994). Remembering can cause forgetting: Retrieval dynamics in long-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20* (5), 1063 – 1087. Retrieved from http://www.apa.org/pubs/journals/xlm/

Barnier, A. J., Hung, L., & Conway, M. (2004). Retrieval-induced forgetting of emotional and unemotional autobiographical memories. *Cognition and Emotion, 18* (4), 457 – 477. doi:10.1080/0269993034000392